From Fixed-event to Fixed-horizon Density Forecasts: Obtaining Measures of Multi-horizon Uncertainty from Survey Density Forecasts

Gergely Ganics[1], Barbara Rossi[2] and Tatevik Sekhposyan[3]

[1]*Banco de España*[*]
[2]*ICREA – Univ. Pompeu Fabra, Barcelona GSE, and CREI*[†]
[3]*Texas A&M University*[‡]

December 10, 2019

Abstract

Surveys of professional forecasters produce precise and timely point forecasts for key macroeconomic variables. However, the accompanying density forecasts are not as widely utilized, and there is no consensus about their quality. This is partly because such surveys are often conducted for "fixed events". For example, in each quarter panelists are asked to forecast output growth and inflation for the current calendar year and the next, implying that the forecast horizon changes with each survey round. The fixed-event nature limits the usefulness of survey density predictions for policymakers and market participants, who often wish to characterize uncertainty a fixed number of periods ahead ("fixed-horizon"). Is it possible to obtain fixed-horizon density forecasts using the available fixed-event ones? We propose a density combination approach that weights fixed-event density forecasts according to a uniformity of the probability integral transform criterion, aiming at obtaining a correctly calibrated fixed-horizon density forecast. Using data from the US Survey of Professional Forecasters, we show that our combination method produces competitive density forecasts relative to widely used alternatives based on historical forecast errors or Bayesian VARs. Thus, our proposed fixed-horizon predictive densities are a new and useful tool for researchers and policy makers.

Keywords: Survey of Professional Forecasters, Density Forecasts, Forecast Combination, Predictive Density, Probability Integral Transform, Uncertainty, Real-time.
JEL codes: C13, C32, C53.

[*]Address: Calle de Alcalá 48, 28014 Madrid, Spain. Tel.: +34-91-338-6135, email: gergely.ganics@bde.es.
[†]Address: C\Ramon Trias Fargas 25-27, Mercè Rodoreda bldg., 08005 Barcelona, Spain. Tel.: +34-93-542-1655, email: barbara.rossi@upf.edu.
[‡]Address: 4228 TAMU, College Station, TX 77843, USA. Tel.: +1-979-862-8857, email: tsekhposyan@tamu.edu.
We thank Todd Clark, Marco Del Negro, Jesús Gonzalo, Michael McCracken, Gabriel Pérez Quirós, Min Chul Shin, and participants of the 1st Vienna Workshop on Economic Forecasting, the VIII$_t$ Zaragoza Workshop in Time Series Econometrics, the 2018 Texas Camp Econometrics, the 2018 Barcelona GSE Summer Forum on Time Series Econometrics and Applications for Macroeconomics and Finance, the 2018 IAAE Annual Conference, the 2018 Conference on Real-Time Data Analysis, Methods, and Applications (Philadelphia Fed), the 2nd "Forecasting at Central Banks" Conference (Bank of England) and seminars at the Federal Reserve Board, Joint Research Centre in Ispra and the Banco de España for comments and useful suggestions. We are grateful to Francesco Ravazzolo and Elmar Mertens for providing codes for the BVAR and the CMM models, respectively. The views expressed herein are those of the authors and should not be attributed to the Banco de España or the Eurosystem. B. Rossi gratefully acknowledges funding from the Fundación BBVA scientific research grant (PR16_DAT_0043) on Analysis of Big Data in Economics and Empirical Applications.

1 Introduction

Central banks around the world aim at improving communication with the public. In line with this effort, there has been an increased emphasis on sharing with the public not only point but also density forecasts, which offer a measure of uncertainty about key macroeconomic variables such as output growth and inflation. For instance, the Bank of England has been publishing fan charts in its Inflation Report since August 1997. In March 2017, the US Federal Open Market Committee started incorporating measures of predictive uncertainty in their Summary of Economic Projections (SEP). The Bank of Canada, the European Central Bank, and the Reserve Bank of Australia are among the many institutions that include information about the uncertainty in their predictions in their monetary policy reports.

Different policy institutions rely on different methodologies to characterize the uncertainty associated with future macroeconomic outcomes, and the best way to construct an accurate density forecast is still unclear. The Bank of England, for instance, uses a mixture of normal distributions to incorporate asymmetry in the predictive density (see Wallis, 1999 for a discussion, and Clements, 2004 and Galbraith and van Norden, 2012 for an evaluation). As discussed in Reifschneider and Tulip (2017), the SEP predictive densities are based on rolling root mean squared forecast errors of the historical point forecasts of private and government forecasters, over a window of twenty years. The SEP densities, unlike the ones provided by the Bank of England, are constructed under the assumption of symmetry. Clark et al. (forthcoming) propose improving the historical-point-forecast-error-based densities by incorporating stochastic volatility in the forecast errors. Rossi and Sekhposyan (2014) and Pettenuzzo et al. (2016), on the other hand, consider fully specified parametric models to construct predictive densities.

Surveys of professional forecasters, in particular, the quarterly US Survey of Professional Forecasters (SPF, currently administered by the Philadelphia Fed) provide precise and timely point forecasts for key macroeconomic variables (see, for example, Ang et al., 2007), justifying their use as a base for the construction of predictive densities as in Clark et al. (forthcoming). Panelists in the US SPF also provide probabilistic forecasts for several variables; for some of them, such as output and prices, since the first survey round of 1968:Q4.[1] However, these predictions are not widely utilized.[2] One of the reasons why density forecasts from the US SPF are not used extensively is because of the format of the survey. In each quarter, survey participants are asked to estimate uncertainty about real GDP growth and inflation (among other variables) for the current and next calendar years. Thus, by construction, the density forecasts are fixed-event forecasts: this means that the forecast horizon changes with survey rounds, limiting their usefulness for policymakers and market participants who instead often seek to characterize uncertainty for a fixed number of periods ahead.[3] In addition, the discrete nature of the histograms and the

[1] Given the changing structure of the survey and the target variables, these density forecasts are typically useful for analysis from 1981:Q3 onward.

[2] Also, there is no consensus about their quality. Zarnowitz and Lambros (1987) were among the first to take a rigorous look at them. Rossi and Sekhposyan (2013) and Rossi and Sekhposyan (2019) formally evaluate their accuracy, while Clements (2014a) and Clements (2018) compare the information in the (consensus) density forecasts to that constructed based on the (unconditional) distribution of point forecasts. These papers provide evidence of misspecification in predictive densities. In addition, the predictive distributions constructed based on the historical forecast errors, on average, are more accurate than the raw densities forecasters provide.

[3] In contrast, this structure is useful for studying the behavior of the forecasters, e.g. how they incorporate newly released information or learn over time. See, for instance, Patton and Timmermann (2010) and Manzan (2017) for the

time-varying structure of the bins used to assign probabilities provide additional complications to researchers and practitioners using probabilistic forecasts. See, for instance, D'Amico and Orphanides (2008), Rossi et al. (2017), Manzan (2017) and Del Negro et al. (2018) for a thorough discussion of how to best fit and conduct inference on professional forecasters' density forecasts provided in the form of histograms.

This paper makes two contributions. First, it proposes a density combination approach to obtain fixed-horizon density forecasts from fixed-event ones. We use an optimal weighting strategy to combine the current and the next year *density* forecasts into a (multi-step-ahead) fixed-horizon density forecast. Several methods have been proposed to transform fixed-event *point* forecasts into fixed-horizon ones, but none have addressed how to do so in the context of density forecasts. For instance, Dovern et al. (2012) use an ad-hoc approach where the weights assigned to the current and the next year point forecasts are proportional to their share of the overlap with the forecast horizon, resulting in deterministic weights. Their method in principle could be applied to densities as well; however, the properties of the resulting density combination are not well known. Knüppel and Vladu (2016), on the other hand, estimate the weights of a linear combination of fixed-event point forecasts with the objective to obtain optimal fixed-horizon point forecasts from a mean squared error perspective. Their approach is not directly applicable to density forecasts. We, instead, propose to estimate the weights with the objective to obtain a correctly calibrated combined predictive density, based on the uniformity of the Probability Integral Transform (PIT) criterion. Our estimator minimizes the distance between the uniform distribution and the empirical distribution of the combined-density-forecast-implied PIT in the Anderson–Darling sense, following Ganics (2017). Our second contribution lies in an extensive investigation of the way to best approximate SPF histograms. Since our density combination methodology requires as input a continuous distribution, we investigate the fit of the normal and skew t distributions to SPF histograms. The resulting combined (fixed-horizon) density is a mixture, thus flexible, possibly featuring asymmetry, multi-modality and fat tails.

We should note that, though obtaining a correctly calibrated combined density is the objective in our benchmark specification, the methodology is not limited to that case. In principle, one can rely on the methodology in Ganics (2017) to target distance measures between any densities (under some regularity conditions). In this context many questions could be of interest. In particular, suppose that the objective of the researcher is to understand how forecasters construct their fixed-event and fixed-horizon density forecasts. For instance, as opposed to the US SPF, the SPF administered by the European Central Bank (ECB-SPF) has information on both. A potentially interesting task would be to recover the weights that make the combined fixed–event density forecasts as close as possible to the reported fixed-horizon ones. In this case the objective function that would pin down the optimal weights is a distance measure between two empirical densities, the combined density and the fixed-horizon one coming from the survey.[4] Alternatively, a practitioner can optimize the weights having some other objective function in mind — for instance, it could be of interest to combine two fixed-event density forecasts such that the resulting combined density maximizes either the log score or continuous ranked probability score (CRPS), two scoring rules typically used for density evaluation.

relevance of point and density forecasts for the study of learning mechanisms.

[4]While we discuss this alternative density forecast estimator in our paper, we do not pursue it since the sample size of the ECB-SPF is too short for a meaningful empirical exercise.

In our paper, instead, we focus on obtaining the "best" calibrated fixed-horizon density forecast. It is important to know that, from a policymaker's point of view, this density might be more useful than the alternatives (for instance, the combined density which maximizes either the log score or the CRPS) since, as shown in Diebold et al. (1998) and Granger and Pesaran (2000), the correctly calibrated density will be preferred by all forecast users, regardless of their loss function. It is also important to clarify that what we mean by "best" calibrated fixed-horizon density forecast is the combination of fixed-event density forecasts whose PIT is the closest to the uniform distribution, and not the density that outperforms certain alternatives according to a particular scoring rule or loss function. For completeness, however, we empirically investigate how our estimated density performs relative to alternative methodologies.

Our results can be summarized as follows. When using the real GDP growth- and GDP deflator-based inflation density forecasts from the US Survey of Professional Forecasters between 1981:Q3 and 2017:Q2, we find that our proposed method indeed delivers correctly calibrated predictive densities for both output growth and inflation (in real time, evaluated based on out-of-sample performance). In terms of relative accuracy of density forecasts, our combination method is competitive against the Bayesian Vector Autoregressive (BVAR) model (with stochastic volatility) recently proposed by Clark and Ravazzolo (2015), densities based on past forecast errors (Clements, 2018), as well as the ones based on Clark et al.'s (forthcoming) stochastic volatility model using nowcast errors and expectational updates. Furthermore, there is little gain from fitting the more flexible skew t distributions to the density forecasts provided by the survey histograms — though the skew t appears to fit better in the Great Recession episode. On the other hand, our combined fixed-horizon-densities are often asymmetric, in particular during the recent financial crisis. This asymmetry is in line with the findings of Adrian et al. (2019) and Manzan (2015), who estimate conditional predictive densities for US real GDP growth using a quantile regression model.

The rest of the paper proceeds as follows. Section 2 lays out the proposed methodology. Section 3 discusses the SPF data and the competing models, while Section 4 presents our results. We conclude with Section 5. Additional robustness checks and results are collected in the Appendix.

2 The Proposed Methodology

In this section, we describe the forecasting environment and introduce the relevant notation. The notation is defined consistently with the structure of the US SPF.[5] However, the framework is not restrictive and could be adapted to any forecasts that share the fixed-event features of the SPF.

2.1 Econometric framework

At each survey round t taking place in quarter $q \in \{1,2,3,4\}$, the survey provides a pair of predictive distributions (cumulative distribution functions or CDFs) corresponding to the variable of interest in the current and the next calendar years, denoted by $\widehat{F}^0_{t,q}(\cdot)$ and $\widehat{F}^1_{t,q}(\cdot)$, respectively.

[5]The ECB's euro area SPF has a richer structure that provides both fixed-horizon and fixed-event probabilistic forecasts. Our notation is also consistent with the fixed-event forecasts of the euro area SPF, keeping in mind the different data release schedules.

We denote the corresponding probability density functions (PDFs) by $\widehat{f}^0_{t,q}(\cdot)$ and $\widehat{f}^1_{t,q}(\cdot)$. Hats ($\widehat{\ }$) indicate that these objects might depend on estimated parameters. Since the target variable for these forecasts does not change as we move throughout the year, i.e. for quarters $q = 1, 2, 3, 4$, these density forecasts are known as fixed-event forecasts in the literature.[6]

We are interested in constructing a density forecast for a variable that is h quarters ahead of the quarter preceding t, whose CDF is denoted by $\widehat{F}^{h,C}_{t,q}(\cdot)$, with the corresponding PDF being $\widehat{f}^{h,C}_{t,q}(\cdot)$.[7] We assume that this CDF (or PDF) is a convex combination of $\widehat{F}^0_{t,q}(\cdot)$ and $\widehat{F}^1_{t,q}(\cdot)$ (or $\widehat{f}^0_{t,q}(\cdot)$ and $\widehat{f}^1_{t,q}(\cdot)$ in the case of the PDF), hence the C superscript. To accommodate the fact that, in each year, there are four quarterly survey rounds with different horizons, the weights are allowed to differ across the quarters in which the survey took place. Let $w^h_q \equiv (w^h_{q,0}, w^h_{q,1})'$ denote the unknown (2×1) weight vector in quarter q for the current and next calendar year forecasts, respectively. The combined predictive distribution we consider is in the class of linear opinion pools, and is given by

$$\widehat{F}^{h,C}_{t,q}(y) \equiv w^h_{q,0}\widehat{F}^0_{t,q}(y) + w^h_{q,1}\widehat{F}^1_{t,q}(y)\,, \tag{1}$$

such that

$$0 \leq w^h_{q,0}, w^h_{q,1} \leq 1\,, \quad w^h_{q,0} + w^h_{q,1} = 1\,, \quad q \in \{1,2,3,4\}\,. \tag{2}$$

The corresponding combined PDF is defined analogously as

$$\widehat{f}^{h,C}_{t,q}(y) = w^h_{q,0}\widehat{f}^0_{t,q}(y) + w^h_{q,1}\widehat{f}^1_{t,q}(y)\,. \tag{3}$$

2.2 Proposed estimator

We propose estimating $\{w^h_{q,0}\}_{q=1}^4$ using the estimator of Ganics (2017), which builds on the fact that a density forecast is probabilistically well-calibrated if and only if the corresponding Probability Integral Transform or PIT (Rosenblatt, 1952; Diebold et al., 1998; Bai, 2003; Corradi and Swanson, 2006; Rossi and Sekhposyan, 2013, 2019) is uniformly distributed.[8] In practice, the weights are estimated by minimizing the distance between the PITs of the combined distribution and the uniform distribution, hence aiming for probabilistic calibration. This requires recording the h-period-ahead realizations of the variable of interest, denoted by $y^h_{t,q}$, and forming the PITs. The PIT is the combined CDF evaluated at the realization, formally:

$$\text{PIT}^h_{t,q} \equiv \widehat{F}^{h,C}_{t,q}(y^h_{t,q}) = w^h_{q,0}\widehat{F}^0_{t,q}(y^h_{t,q}) + w^h_{q,1}\widehat{F}^1_{t,q}(y^h_{t,q}) = \int_{-\infty}^{y^h_{t,q}} \widehat{f}^{h,C}_{t,q}(y)\,\mathrm{d}y\,. \tag{4}$$

[6]The US SPF provides the users with histograms. Some papers in the literature have chosen to fit a PDF over the histograms, while others have taken the route of fitting a CDF over cumulative histograms. The latter is the approach of the current paper since by cumulating the histograms, we get directly the CDF, while we would need further assumptions on where the probability mass is (say, at the midpoint of the histogram bin, for example) if we were to fit the PDF on the histograms.

[7]Our empirical implementation takes into account the fact that in survey round t, panelists have access to data of the previous but not the current quarter since we are interested in real-time properties of the combined density. The framework could be adapted to alternative timing assumptions.

[8]Note that, as the weights in a given quarter sum to one ($w^h_{q,0} + w^h_{q,1} = 1$), it is sufficient to estimate the weights associated with current year's forecasts ($w^h_{q,0}$).

We should note that in the empirical application we use h-quarter-ahead (from the quarter preceding t) year-on-year growth rates for the quarterly economic variables of interest. However, the framework is general, and could be applied to any definition of a realization. In other words, the realization does not have to be consistent with the definition of the target in the input densities — the density combination weights serve as an adjustment device for obtaining an optimally calibrated fixed-horizon-density forecast for the target definition of researcher's choice.[9]

Consider the vertical difference between the empirical distribution function of the PITs and the CDF of the uniform distribution at quantile $r \in [0,1]$:

$$\Psi_{\mathcal{T}}(r, w_q^h) \equiv |\mathcal{T}|^{-1} \sum_{t \in \mathcal{T}} \mathbb{1}\left[\text{PIT}_{t,q}^h \leq r\right] - r\,, \tag{5}$$

where $\mathcal{T}$ is the index set of an appropriate sample of size $|\mathcal{T}|$ and $\mathbb{1}[\cdot]$ is the indicator function. One might take $\mathcal{T}$ as all the years in a sample for which there is an observed realization of $y_{t,q}^h$ and estimate the weights separately for each quarter. For example, if the sample starts in the third quarter, this would mean taking every fourth observation, starting with the first one, to estimate the weight w_3^h corresponding to the third quarter surveys.[10] In a sample of 100 consecutive quarterly observations, this would mean using only 25 observations ($\mathcal{T} = 1, 5, \ldots, 97$, $|\mathcal{T}| = 25$) to estimate each quarter-specific weight, possibly resulting in considerable estimation uncertainty and excluding the possibility of producing out-of-sample forecasts using the estimated weights.

However, to accommodate the small sample sizes often encountered in practice, and to perform out-of-sample forecasting exercises, we parametrize the weights as flexible, exponential Almon lag polynomials (Andreou et al., 2010):

$$w_{q,0}^h \equiv \exp(\theta_1 q + \theta_2 q^2)\,, \quad q \in \{1,2,3,4\}\,, \tag{6}$$

$$w_{q,1}^h \equiv 1 - w_{q,0}^h\,, \tag{7}$$

$$w^h \equiv (w_{1,0}^h, w_{2,0}^h, w_{3,0}^h, w_{4,0}^h)'\,, \tag{8}$$

and adopt a rolling window estimation scheme by taking $\mathcal{T} = s - R + 1, s - R + 2, \ldots, s$, where $s = R, R+1, \ldots, T$ is the last observation of a rolling window of size R, and T is the last available density forecast observation in the full sample. The parametrization in Equation (6) guarantees that the weights are positive and allows us to pool together PITs from different quarters, using an exponential polynomial.[11] Accordingly, we estimate weights via a modified version of the Anderson–Darling-type weight estimator of Ganics (2017), which is defined as the minimizer of

[9] For instance, the surveys could be asking for quarter-on-quarter growth rates, yet the researcher could be interested in obtaining the "best" combination for the year-on-year growth rate, in which case the appropriate measure of realization for the construction of the PITs would be the year-on-year growth rate.

[10] Empirical results on this exercise are available upon request and are used as a motivation for a parametric specification of the weight function such that we can conduct a truly out-of-sample density combination exercise.

[11] There are other types of polynomials that could help us obtain positive weights while being parsimonious. For instance, Ghysels et al. (2007) propose the use of a beta function. The Almon lag polynomial remains a popular choice in the mixed data sampling (MIDAS) literature, which we borrow from directly, though the role of the polynomial in our context is to provide weights for density combination as opposed to providing weights for high frequency data aggregation (as would be in the MIDAS case). Ghysels et al. (2007) do not find dramatic differences in the performance of the polynomials in the MIDAS framework, if anything, they recommend the Almon lag polynomial for data at lower frequency (which is the case that we consider) and the beta polynomial for data at higher frequency.

the scaled quadratic distance:

$$\widehat{w}^h_{q,0} \equiv \exp(\widehat{\theta}_1 q + \widehat{\theta}_2 q^2)\,, \quad q \in \{1,2,3,4\}\,, \tag{9}$$

$$(\widehat{\theta}_1, \widehat{\theta}_2)' \equiv \underset{\theta_1,\theta_2\in\Theta}{\operatorname{argmin}} \int_\rho \frac{\Psi^2_{\mathcal{T}}(r, w^h)}{r(1-r)}\,\mathrm{d}r\,, \tag{10}$$

where the parameter space Θ is set to ensure that the weights satisfy $0 < \widehat{w}^h_{q,0} \leq 1$ for $q \in \{1,2,3,4\}$, and they are non-increasing in q.[12] The motivation for the latter restriction is that, intuitively, as we progress through the year from quarter q to $(q+1)$, we do not wish to give more weight to current year's forecast in $(q+1)$ than we did in q.[13] Finally, $\rho \subset [0,1]$ is a finite union of neither empty nor singleton, closed intervals on the unit interval, where we wish to minimize the Anderson–Darling-type discrepancy between the empirical CDF of the PIT and the uniform CDF.[14] In our empirical application, we will use $\rho = [0,1]$, and the minimization is implemented numerically (via MATLAB's `fmincon` function).

Ganics (2017) proves the consistency of the weights obtained by a similar minimization problem, where the weights are estimated directly, without the exponential Almon lag parametrization under some regularity and identification conditions — the former ensure that a weak law of large numbers holds for the empirical CDF of the PITs and that the PIT is continuously distributed, while the latter guarantees the uniqueness of the true weight vector. The exponential Almon lag parametrization requires that the identification condition holds in the parameter space Θ instead of the unit simplex where the weights themselves are. It is important to notice that the weight estimator does *not* require the individual $\widehat{F}^0_{t,q}(y)$ and $\widehat{F}^1_{t,q}(y)$ distributions to be correctly calibrated, but rather the combination procedure itself serves as a device to achieve the best correctly calibrated combined distribution.

2.3 Alternative fixed-horizon estimators

An alternative method for assigning weights to fixed-event *point* forecasts to obtain one-year-ahead (fixed-horizon) *point* forecasts is proposed by Dovern et al. (2012). They suggest using ad-hoc deterministic weights proportional to the share of the overlap that fixed-event forecasts have with the chosen fixed forecast horizon. Dovern et al. (2012) combine fixed-event point forecasts available at a monthly frequency. Rossi et al. (2017) consider an extension of this method to

[12] In particular, in order to make the weights non-increasing in q, we restrict the domain of θ_1 and θ_2 by imposing the following constraints. Let $K = \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \\ 1 & 7 \end{pmatrix}$, $b = (0,0,0,0)'$. The weights are positive due to the exponential function. Imposing the constraint $K\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \leq b$ ensures $\widehat{w}^h_{1,0} \leq 1$ (thanks to the first inequality of the constraint), and that $\widehat{w}^h_{1,0} \geq \widehat{w}^h_{2,0} \geq \widehat{w}^h_{3,0} \geq \widehat{w}^h_{4,0}$ (implied by the second to fourth inequalities of the constraint).

[13] In an earlier version of the paper, we estimated the weights (in full-sample) quarter by quarter, and indeed found a non-increasing pattern.

[14] The reason why we use the Anderson–Darling objective function is because Ganics (2017) shows it performs better than some of the alternatives such as Cramér–von Mises- or Kolmogorov–Smirnov-type objective functions in Monte Carlo simulations in terms of overall out-of-sample density calibration of the implied density combination. Relative to these alternatives, which could be easily applied in our empirical context, the Anderson–Darling measure of discrepancy between the empirical and theoretical CDFs has more penalty associated with the tails. For further details on the estimator, see Ganics (2017).

one-year-ahead ($h = 4$) quarterly *density* forecasts. Based on this approach, in quarter one, the current year density gets a weight of one and the next year density a weight of zero, while in quarter two those weights would be three fourth and one fourth, respectively, and so on. Formally, in a particular quarter q the weights are

$$w^4_{q,0} = \frac{5-q}{4}, \qquad w^4_{q,1} = \frac{q-1}{4}, \quad q \in \{1,2,3,4\}. \tag{11}$$

One advantage of Dovern et al.'s (2012) method is that the weights are given and need not be estimated, which is useful in light of the small sample sizes often encountered in practice. However, it is unclear whether an analogous combination scheme is applicable in situations with different forecast horizons or frequencies.[15] Furthermore, these weights are intended to be the same for any density combination and do not take into account the specific features of the data generating process or the input densities. As mentioned earlier, Knüppel and Vladu's (2016) estimator overcomes this problem, but it is specifically designed for *point* forecasts only.

2.3.1 Fixed-horizon predictive densities that mimic forecasters' predictions

As noted earlier, we aim at obtaining correctly calibrated density forecasts, as they will be preferred by all forecasters among a set of alternatives, regardless of their specific loss functions (Diebold et al., 1998 and Granger and Pesaran, 2000). To achieve this objective, we choose to minimize the distance between the CDF of the PIT of the combined density and that of a uniform, i.e. a 45° line, represented by r in Equation (5).

In principle, a researcher can choose to minimize the distance between the CDF of the combined distribution and some other distribution. This choice of the alternative density, in itself, depends on the question of interest. For instance, a researcher can choose to extract one-step-ahead forecast information from the current-year and next-year forecasts in each quarter. The information on the one-quarter-ahead forecasts is embedded in the current-year and next-year forecasts; however, the weight associated with that informational component is unknown. One of the ways a researcher can address this issue is to consider a density combination approach, where the current-year and next-year densities are weighted to match the one-quarter-ahead density forecasts provided by SPF participants at least once a calendar year (say the current year forecasts in the fourth quarter of the year). More formally, a researcher can choose to minimize the following distance measure

$$\underset{w^1_q}{\operatorname{argmin}} \int_\rho \frac{\widetilde{\Psi}^2_{\mathcal{T}}(r, w^1_q)}{r(1-r)}\, \mathrm{d}r, \text{ for } q = 1,2,3, \tag{12}$$

where $\widetilde{\Psi}_{\mathcal{T}}(r, w^1_q) \equiv |\mathcal{T}|^{-1} \sum_{t\in\mathcal{T}} \left[\widehat{F}^{1,C}_{t,q}(y^1_{t,q}) - \widehat{F}^0_{t,4}(y^1_{t,4})\right]$. Naturally, the minimization process would not be required for $q = 4$, since in that quarter the density $\widehat{F}^0_{t,4}(y^1_{t,q})$ would be taken as given. This is one example on how one could modify the proposed density weighting strategy to extract information from fixed-event-forecasts — here, we obtain one-quarter-ahead densities, and there

[15] For instance, it is not obvious how the ad-hoc weights would look like for, say, a five-quarter-ahead forecast horizon. For example, it is unclear if the researcher would discard the current year forecasts in the density combination approach and, if not, what weight he or she would assign to it.

are no optimality conditions imposed. These densities could potentially be misspecified since they are only asked to emulate the properties of the current year fixed-horizon forecast in the fourth quarter of the year (i.e. one-step-ahead forecasts).

3 An Overview of the Data and Models

In this section, we discuss how to obtain a fixed-horizon predictive density for the SPF dataset in real-time. We then outline alternative approaches that have been shown to produce competitive density forecasts. Some of these models rely on the point forecasts provided by the professional forecasters.

3.1 The data

We construct four-quarter-ahead density forecasts of quarterly year-on-year US real GDP growth and inflation measured by the GDP deflator, based on the US SPF. In what follows, we will refer to these variables as GDP growth and inflation, respectively.

For both variables, we use SPF surveys between 1981:Q3 and 2017:Q2. In the US SPF, panelists are asked to provide their probabilistic forecasts of the growth rate of the average level of real GDP and the GDP deflator from the previous calendar year to the current calendar year, and from the current calendar year to the next calendar year. We choose the beginning of the sample period in order to obtain the longest possible sample for (approximately) the same key variables. In particular, the SPF documentation (Federal Reserve Bank of Philadelphia, 2017) states, "*The old version (prior to 1981:Q3) asked for the probability attached to each of 15 possible percent changes in nominal GNP and the implicit deflator, usually from the previous year to the current year. The new version (1981:Q3 on) asks for percent changes in real GNP and the implicit deflator, usually for the current and following year.*" It also asserts that "*Then, in 1992:Q1, the number of categories was changed to 10 for each of the two years, and output was changed from GNP to GDP.*" To mitigate the effect of these changes, our sample starts in 1981:Q3.

The probabilistic forecasts in the SPF take the form of probabilities assigned to pre-specified bins, and we must transform these into a continuous PDF prior to the analysis in order to satisfy the continuity assumption for the weight estimator. In what follows, we describe the procedure to obtain these continuous density forecasts.[16] To simplify the notation, we suppress time indices t and q.

[16] An alternative to fitting the continuous distributions over the provided histograms is to use probability integral transforms for count data as suggested by Czado et al. (2009) and Kheifets and Velasco (2017). For discrete support random variables, which the SFP histograms indeed are, the PIT is no longer uniform under probabilistic calibration. As discussed in the aforementioned papers, the literature has relied on interpolation methods, where independent random noise is introduced either to the data or to the PITs to recover uniformity. The performance of these types of methods is not ideal since the additional noise can distort some of the outcomes. Czado et al. (2009), on the other hand, consider a non-randomized version of the PITs, which behaves as a uniform random variable under probabilistic calibration. Further, Kheifets and Velasco (2017) provides a framework for testing for proper calibration for the non-randomized version of the PITs constructed on the count data. We could indeed use this version of the PIT in our empirical application, since this would satisfy the continuity assumption under which the weight estimator operates. We chose not to do that in the current context for two reasons. First, all of the empirical literature we are aware of that looks at the SPF density forecasts uses some version of continuous approximation to the histograms. Second, we would need further simulations studies to see how the construction of the PITs under the count data assumption performs relative to the case where the histograms are approximated with a continuous distribution. We leave this question to future research.

Using the average of individual survey respondents' predictions for each bin, in each quarter (survey round) we calculate the empirical CDF for the GDP growth and inflation forecasts separately, for both the current year and the next. In our analysis, by taking the average of individual survey responses, we form a "consensus" forecast, similarly to how the Philadelphia Fed formulates the consensus forecast, and we do not investigate individual panelists' beliefs.[17] For a Bayesian approach focusing on that problem, see Del Negro et al. (2018). Formally, in a given quarter, let $\{s_i\}_{i=1}^{S}$ denote the set of endpoints associated with the intervals/bins specified by the survey and let $F(s_i)$ denote the value of the empirical CDF implied by the SPF histogram at s_i.[18] This set contains the right endpoints of all the bins except the last bin, whose left endpoint is the only one included. The survey is designed such that the leftmost (rightmost) bin is open from the left (right), and we do not impose an arbitrary s_0 or s_{S+1}, where $F(s_0) = 0$ and $F(s_{S+1}) = 1$.

3.2 The models

We fit both a normal and Jones and Faddy's (2003) skew t distribution (hereinafter JF distribution) to the resulting CDF. The normal distribution is frequently used in the literature[19] (see e.g. Giordani and Söderlind, 2003, D'Amico and Orphanides, 2008, Clements, 2014b or Rossi et al., 2017); to the best of our knowledge, we are the first to use the aforementioned skew t distribution to model survey forecasts. This is motivated by the observation that, as we will illustrate later on, in a number of cases, the survey forecast histograms seem better represented by a skewed underlying distribution. Jones and Faddy's (2003) skew t distribution generalizes Student's t distribution by introducing the parameters $a, b > 0$ regulating skewness and tail behavior at the same time. It is important to note that Jones and Faddy's (2003) distribution encompasses as special cases both the usual Student's t (when $a = b$) and the normal distribution (when $a, b \to \infty$). After introducing a location and a scale parameter, denoted by μ and $\sigma > 0$ respectively, the distribution's PDF at $x \in \mathbb{R}$ is given by

$$f(x; \mu, \sigma, a, b) = \frac{1}{\sigma} C_{a,b}^{-1} (1 + \tau)^{a+1/2} (1 - \tau)^{b+1/2} , \tag{13}$$

$$C_{a,b} = 2^{a+b-1} B(a, b)(a + b)^{\frac{1}{2}} , \tag{14}$$

$$\tau = \frac{x - \mu}{\sigma} \left(a + b + \left(\frac{x - \mu}{\sigma} \right)^2 \right)^{-\frac{1}{2}} , \tag{15}$$

[17] In principle, our proposed methodology could be applied to individual panelists' forecasts, provided we can obtain a panel of forecasters that provide fixed-event density forecasts for both current and next calendar years. Furthermore, it could be used as an optimal method of aggregation across the forecasters.

[18] In the literature, some papers fit the moments, while others fit the cumulative distribution function, see Clements (2014b, p. 101) for a discussion. We follow a procedure similar to that of Engelberg et al. (2009).

[19] An interesting alternative is the generalized beta/triangular distribution used by Engelberg et al. (2009) and Clements (2014b). When using a beta distribution, one would need to close the open tail bins of the SPF histograms. Using a normal and skew t distribution has the advantage that the tail probabilities are well defined, thus there is no need to close the tail bins of the histograms in an arbitrary manner.

while its CDF is given by

$$F(x;\mu,\sigma,a,b) = I_z(a,b)\,, \tag{16}$$

$$z = \frac{1}{2}\left(1 + \frac{\left(\frac{x-\mu}{\sigma}\right)}{\sqrt{a+b+\left(\frac{x-\mu}{\sigma}\right)^2}}\right)\,, \tag{17}$$

where $B(\cdot,\cdot)$ is the beta function and $I_v(\cdot,\cdot)$ is the regularized incomplete beta function (also known as the incomplete beta function ratio).[20]

Let $d = \{\text{N}, \text{JF}\}$ index the normal and the JF distributions, respectively. Let θ collect the parameters of either the normal distribution, corresponding to $\theta_{\text{N}} = (\mu, \sigma^2)'$, or the JF distribution, where $\theta_{\text{JF}} = (\mu, \sigma, a, b)'$. In the former case, the parameter space is $\Theta_{\text{N}} = \mathbb{R} \times \mathbb{R}^+$, while in the latter case we restrict the skewness parameters to $a, b > 2$ to ensure the existence of the first four moments, implying $\Theta_{\text{JF}} = \mathbb{R} \times \mathbb{R}^+ \times (2,\infty) \times (2,\infty)$. The parameters of each distribution are estimated using non-linear least squares, given the set of endpoints $\{s_i\}_{i=1}^S$ associated with the intervals/bins specified by the survey:

$$\widehat{\theta}_d = \underset{\theta_d \in \Theta_d}{\operatorname{argmin}} \sum_{i=1}^{S} \left(F_d(s_i;\theta_d) - F(s_i)\right)^2\,. \tag{18}$$

By estimating the distributions in each quarter, the procedure described above gives us sequences of predictive CDFs $\widehat{F}^0_{t,q,d}(y)$ and $\widehat{F}^1_{t,q,d}(y)$ for each variable of interest. Analogously, the corresponding predictive PDFs are denoted by $\widehat{f}^0_{t,q,d}(y)$ and $\widehat{f}^1_{t,q,d}(y)$. For example, take the 2009:Q2 survey round and consider forecasts of GDP growth. Figure 1 shows the empirical CDFs and histograms for the current year and next year GDP growth, together with the fitted normal and skew t CDFs and PDFs (skew t denoted by ST^{JF}).[21] The fitted distributions range from the 1981:Q3 survey round to the 2017:Q2 round, but we did not use the 1985:Q1 and 1986:Q1 surveys due to an error (documented in Federal Reserve Bank of Philadelphia 2017, p. 25). Hence, the full sample contains 142 quarterly surveys.

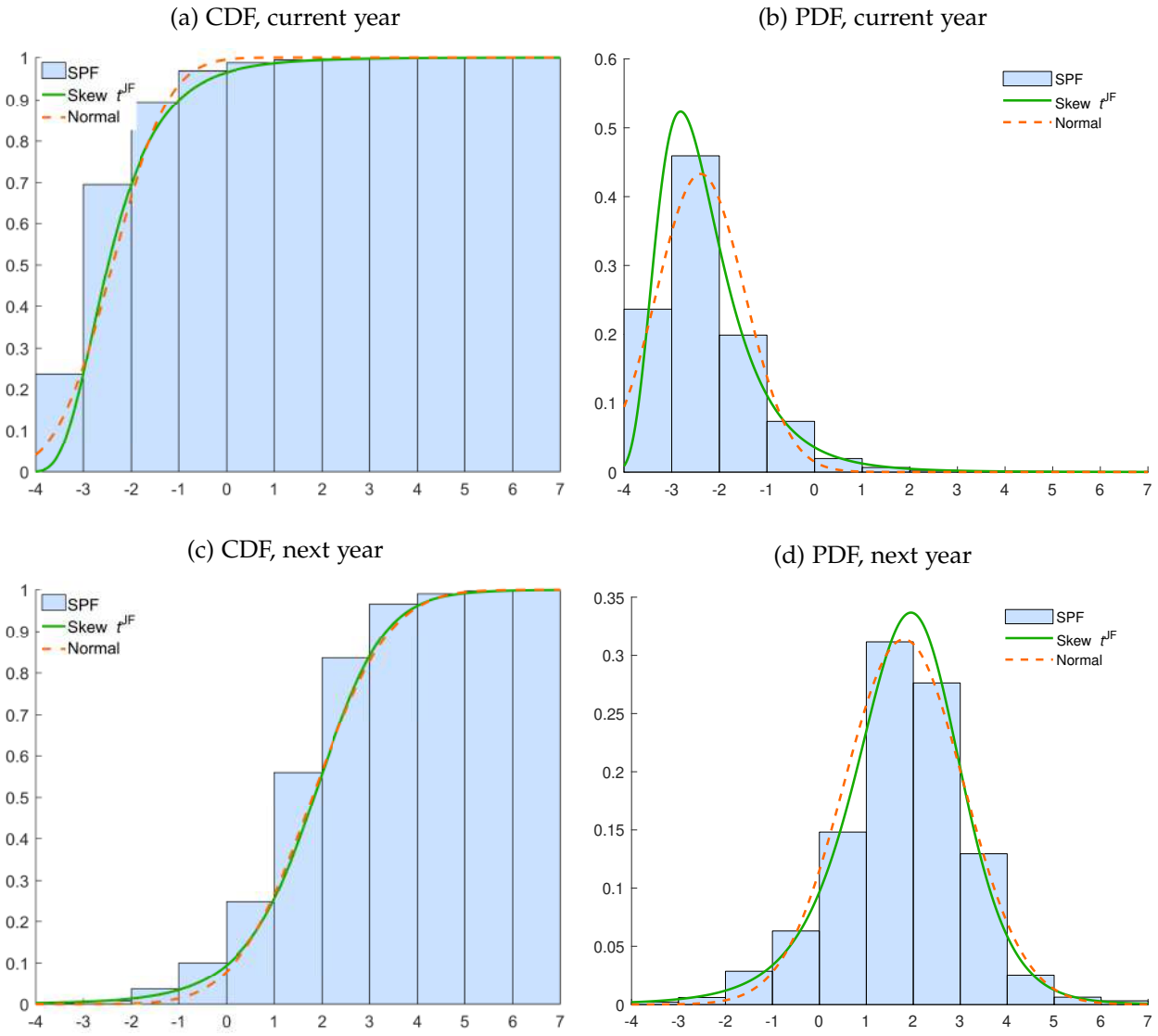
Figure 2 shows the skewness (standardized third central moment) of the fitted distributions between the 1981:Q3 and 2017:Q2 survey rounds. It is interesting to note that current year's forecasts are usually more asymmetric than next year's forecasts; furthermore, GDP growth forecasts are mostly negatively skewed, while inflation forecasts are usually positively skewed.

To form the sequence of PITs, we used GDP growth and GDP deflator data (both seasonally adjusted) from the Federal Reserve Bank of Philadelphia's Real-Time Data Research Center. In line with the information set of the survey respondents, in any given quarter, the latest measurement of the variable of interest that the forecaster has access to corresponds to the *previous* quarter,

[20] There are alternative ways to construct skew t distributions, such as Azzalini and Capitanio's (2003) skew t, which was recently used by Adrian et al. (2019). In order to ensure that our results are not driven by our specific choice, we also performed a robustness analysis using Azzalini and Capitanio's (2003) distribution. As Appendix A demonstrates, our results are largely unchanged. This is due to the fact that the two distributions fit the data in a similar way (see Azzalini and Capitanio 2014, pp. 105–108). However, the Jones and Faddy (2003) distribution is particularly appealing due to the fact that its CDF can be evaluated very quickly, considerably speeding up the estimation procedure since the incomplete beta function is readily available in several econometrics packages (such as MATLAB). The CDF of Azzalini and Capitanio's (2003) skew t requires numerical integration, hence more computation time. For brevity, the formulas for the PDF and the CDF of the normal distribution are omitted.

[21] We fit the CDFs and then plot the implied PDFs.

Figure 1: Fitting normal and skew t CDFs to the SPF empirical CDFs of GDP growth in 2009:Q2



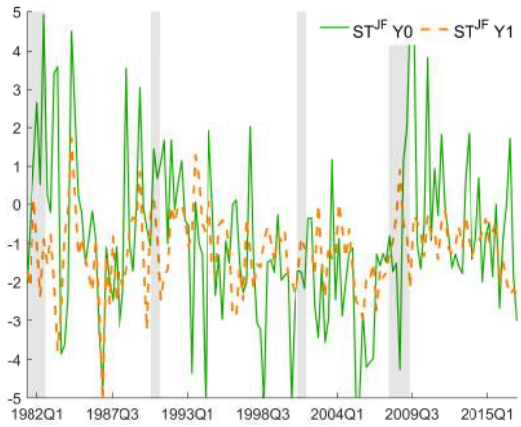
Note: The figures show the empirical CDFs and histograms of the SPF forecasts of GDP growth for 2009 and 2010, according to the 2009:Q2 survey round, and the fitted skew t (solid green curve) and normal (dashed orange curve) distributions' CDFs and PDFs.

which we take as the point of reference. Therefore, by four-quarter-ahead forecasts we mean four quarters after the quarter preceding the survey round ($h = 4$ in the notation of Section 2). The four-quarter-ahead realizations are calculated based on the first/advance estimates of GDP and the GDP deflator. For example, the first GDP growth realization in our sample (corresponding to the 1981:Q3 survey round) is constructed as follows. The quarter preceding the survey is 1981:Q2, while the date four quarters later is 1982:Q2. The first estimate of GDP corresponding to 1982:Q2 was published in 1982:Q3. The percentage growth rate of the GDP estimates in 1982:Q2 and 1981:Q2 according to the 1982:Q3 vintage gives us the first realization of real-time GDP growth. Thus, the full sample ranges from 1982:Q2 to 2018:Q1.

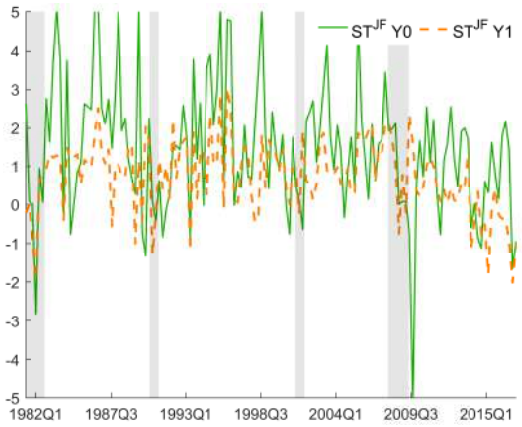
We estimate the combination weights in rolling windows of $R = 60$ quarters using the Anderson–Darling-type objective function in Equation (10), with $\rho = [0, 1]$. For example, in the case of GDP growth, the first pseudo out-of-sample prediction, corresponding to 1998:Q3, is obtained as follows. To estimate the combination weights in Equation (9), the first rolling

Figure 2: Skewness of fitted skew t distributions between 1981:Q3 and 2017:Q2

(a) GDP growth



(b) Inflation



Note: Dates correspond to US SPF survey rounds. Solid green lines (dashed orange lines) show the skewness of the fitted JF skew t distributions corresponding to current year's (next year's) GDP growth or inflation. Shaded areas denote NBER recession periods.

window uses the fitted distributions (either normal or skew t) between the 1981:Q3 and 1996:Q4 survey rounds, and the corresponding GDP growth realizations between 1982:Q2 and 1997:Q3. These weights would have been available in the 1997:Q4 survey round at the earliest (due to the publication delay of the 1997:Q3 GDP data), therefore we combine the fixed-event predictive distributions in this survey using the combination weights $(\widehat{w}^4_{4,0}, \widehat{w}^4_{4,1})'$.[22] In the next rolling window, we re-estimate the weights using survey data between 1981:Q4 and 1997:Q1 and realizations between 1982:Q3 and 1997:Q4 and apply the weight estimates $(\widehat{w}^4_{1,0}, \widehat{w}^4_{1,1})'$ to the fixed-event surveys of 1998:Q1, resulting in a GDP growth forecast for 1998:Q4. This procedure is repeated until the end of the sample, with survey rounds between 2003:Q3 and 2016:Q2, GDP growth data between 2004:Q2 and 2017:Q1, and the last weights being applied to the 2017:Q2 survey forecasts, predicting GDP growth for 2018:Q1. As a result, we obtain fixed-horizon density estimates in real-time, i.e. only using the information that was available at the forecast origin date.

In the next section, the models using the normal or the skew t distribution are denoted by N or ST$^{\text{JF}}$, respectively. We compare our suggested weight estimation procedure against the ad-hoc, fixed weights given in Equation (11). When the ad-hoc mixture weights are used, we added "(ah)" after the distribution's abbreviation, while the models using estimated weights are not accompanied by additional notation. It should be noted that both the optimal as well as the ad-hoc weights generate combined densities which are mixtures of either normal or skew t distributions and, thus, are potentially flexible and have the ability to showcase multi-modality, asymmetry, fat tails, etc. As discussed earlier, the ad-hoc weights come with the advantage that they do not need to be estimated, the disadvantage being that it is not obvious how to derive ad-hoc weights for general cases (see Footnote 15).

[22] For simplicity we did not label the weight estimates according to the estimation sample.

3.3 Benchmark models

To illustrate the merits of the proposed density combination procedure, we compare it to alternative approaches typically used in the literature, which we describe in this section. It is important to note at the onset that our objective is to obtain a combination of fixed-event density forecasts whose PIT is the closest to the uniform distribution, and not the density that outperforms certain alternatives according to a particular scoring rule or loss function, although, for completeness, we empirically investigate how our estimated density performs relative to the alternative methodologies.

3.3.1 BVAR with stochastic volatility

The first benchmark model is based on Clark and Ravazzolo (2015), who show that a BVAR model with stochastic volatility provides competitive density forecasts. Our BVAR specification corresponds to the "VAR-SV" model in the aforementioned paper, and includes: GDP (100 times logarithmic difference), the GDP deflator (100 times logarithmic difference), the unemployment rate (quarterly average of monthly data) and the 3-month Treasury bill rate (quarterly average of monthly data). The BVAR has 4 lags. In order to account for potential parameter instability but keep the computational costs manageable, the model is re-estimated in rolling windows of 60 quarterly observations (15 years), using real-time data for the GDP and GDP deflator (the Treasury rate series is not subject to revisions, while revisions to unemployment rate are minor, hence we followed Clark and Ravazzolo, 2015 and used the latest vintage). The raw GDP and GDP deflator data are obtained as before, while the 3-month Treasury bill rate series is downloaded from the website of the Federal Reserve Board of Governors (H.15 Selected Interest Rates, series H15/H15/RIFSGFSM03_N.M), and the unemployment rate is downloaded from the FRED database maintained by the Federal Reserve Bank of St. Louis (mnemonic: UNRATE).

In order to obtain real-time forecasts, only data available up to the previous quarter are used in the estimation procedure at each forecast origin. In line with the forecast combination scheme described earlier, the BVAR estimated in the first rolling window provides four-quarter-ahead GDP growth and inflation forecasts for 1998:Q3. We use the same prior specification as Clark and Ravazzolo (2015), except that the parameters used in the prior distribution (which depend on a training sample of 24 pre-sample observations) are re-set in each rolling window. After a burn-in period of 20,000 draws, we simulate 80,000 draws from the BVAR's four-quarter-ahead predictive density and retain one out of 8 draws. In the BVAR model, GDP and GDP deflator enter as quarterly growth rates. We transform the draws from the predictive distribution to obtain year-on-year growth rates for these variables.

3.3.2 The PFE model

As a second benchmark model, we estimate predictive distributions based on past forecast errors (PFE) using the SPF point forecasts. The good forecasting performance of this approach has recently been demonstrated by Clements (2018), in particular for annual average output growth at the one-quarter-ahead horizon, and annual average inflation from one to three quarters ahead. However, unlike Clements (2018), we are interested in fixed-horizon rather than fixed-event forecasts.

In constructing the forecast density, we do not model time-varying volatility directly, but rather let a rolling window estimation scheme account for changes in the variance of the predictive distribution. The predictive density is Gaussian and obtained as follows. At each survey round, we calculate four-quarter-ahead year-on-year forecast errors using the past 60 forecast–observation pairs. For example, in the case of GDP growth, in the 1997:Q4 survey round, we take the 1997:Q4 vintage of GDP (in levels) and calculate the year-on-year growth rate of GDP between 1982:Q4 and 1997:Q3. Then, we take the four-quarter-ahead year-on-year GDP growth point forecasts between the 1982:Q1 and 1996:Q4 surveys.[23] Next, we estimate the mean squared forecast errors and set the variance of the predictive distribution equal to this estimate. The mean of the predictive distribution, on the other hand, is obtained as the four-quarter-ahead point forecast of the actual 1997:Q4 survey. In fact, Clark et al. (forthcoming) use a similar benchmark and show that it performs well, particularly in the sample period considered in our analysis.

3.3.3 The CMM model

Our final benchmark model is the one recently proposed by Clark et al. (forthcoming), who use the historical forecast errors to obtain multi-step-ahead density forecasts. They do so by modeling the multi-step-ahead forecast errors as a sum of the nowcast error (based on the first-release data, as is in our case) and expectational updates (that is, the changes in the point forecast for the same target variable from one survey round to the next) extracted from the US SPF. The model — henceforth referred to as CMM — is estimated using Bayesian Markov Chain Monte Carlo (MCMC) methods in a rolling manner, using 60 quarterly observations. We focus on their baseline specification, which assumes a martingale difference property for the expectational updates, where the innovations are random variables with stochastic volatility. In our exercise the first estimation window is chosen such that the forecast periods in their approach aligns with ours. As their model generates one-quarter-ahead predictive distributions of the *forecast errors*, we first simulate the time-path of the forecast errors four quarters ahead, then obtain draws for the *forecasts* by adding them to the point forecasts readily available from the SPF (as suggested in the description of the forecasting algorithm on pp. 16-17 of Clark et al., forthcoming), and finally convert these into quarterly year-on-year forecasts of GDP growth and inflation. At each forecast origin, after discarding a burn-in of 10,000 draws, we store every 100th draw from the MCMC output, resulting in 10,000 draws used in our analysis. For further details on the model and the estimation procedure, see Clark et al. (forthcoming).

3.3.4 Relationship between our approach and the benchmark approaches

It is important to note that the alternative approaches imply unimodal and symmetric predictive distributions for the quarter-on-quarter growth rates of the variables of interest, at least in large samples. In the case of the BVAR, departures from unimodality and symmetry could be observed due to parameter estimation error in small samples as well as from transforming the original quarter-on-quarter growth rate forecasts into year-on-year ones. On the other hand, the CMM and PFE predictive distributions are unimodal and symmetric by construction, while our mixture

[23] The growth rate forecasts are constructed from the levels point forecasts of the SPF in order to adhere to the year-on-year definition of the target variable.

densities can results in multi-modality and asymmetry due to the distributional properties of the component distributions as opposed to parameter estimation error.[24]

4 Empirical Results

In this section, we present the empirical results of our density combination method.

Figure 3 shows the time series of the estimated weights using the mixture of normal distributions and the mixture of skew t distributions over time for each variable. Recall that in each SPF round, Equations (9) and (10) provide weight estimates $\widehat{w}_{q,0}^4$ for $q = 1, \ldots, 4$, which is what we display in Figure 3. However, when we combine the densities, we only use the value which corresponds to the quarter of the particular SPF round. In the case of GDP growth, we can see in Panels a and c that the estimated weights display substantial time-variation. Panels b and d, instead, display an entirely different pattern for inflation, where the current year's density forecast receives almost all the weight at all time periods, with minor exceptions in the case where the underlying histograms are approximated with a normal distribution.[25]

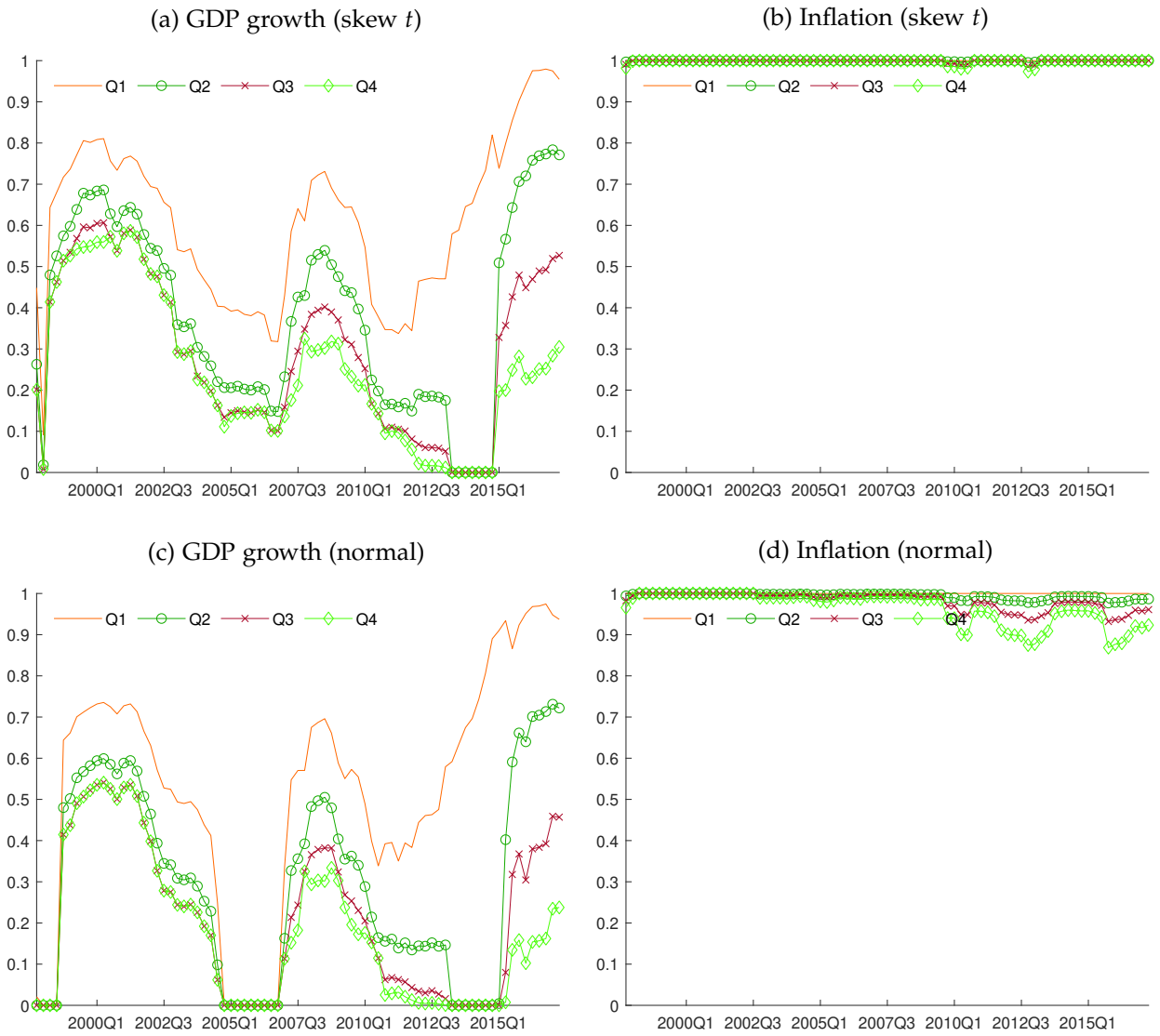
Panels a, c, and e of Figure 4 show the mean, standard deviation and skewness of the combined fixed-horizon predictive densities and the benchmark models for GDP growth, while Panels b, d, and f display the same for inflation forecasts. We can see that the mean of the survey-based forecasts accurately trace the realizations for both variables. In particular, it is interesting to see that the BVAR consistently underpredicted inflation before the Great Recession, and overpredicted it afterwards. In terms of standard deviations, in the case of GDP growth the PFE provides by far the most dispersed predictive distribution, usually followed by the BVAR model, while for inflation the BVAR model stands out with a high standard deviation for only a short period after the recent recession. For GDP growth, the CMM model usually has one of the lowest standard deviations (apart from a period following the recession in the early 2000s), and its predictive distribution became more dispersed during the Great Recession. For inflation, its standard deviation is often lower than that of the combination methods before the recent recession, but after the crisis the CMM model displays similar standard deviations. As for asymmetry, the mixtures of normal distributions (which *could* be skewed in theory) display very little skewness for both GDP growth and inflation apart from a few periods. However, the mixture of skew t distributions shows a markedly different pattern: GDP growth forecasts are considerably negatively skewed, while inflation forecasts are most often strongly positively skewed. The CMM model implies symmetric predictive distributions, as mentioned earlier. The BVAR, on the other hand, displays very little skewness, except in the case of the GDP growth towards the end of the sample period.

Figure 5 shows the predictive distributions (as opposed to the first three central moments) in 2009:Q2 as well as the corresponding target realization of the GDP growth. It is interesting to see that the mixture densities exhibit strong bimodality, with one mode being very close to the actual

[24] The transformation from quarter-on-quarter growth rates to year-on-year ones could result in departures from symmetry in the case of the CMM model, although empirically this does not seem to be relevant, see Section 4.

[25] We have investigated whether this behavior is due to the lack of identification of the weights. As it turns out, in general, the objective function is not flat in the neighborhood of the estimated parameters, suggesting at least local identification (see an example in Figure B.2). Furthermore, Figure B.1 suggests that, for inflation, the component densities associated with the current year forecasts are, on average, better calibrated. This evidence could rationalize putting more weight on the current year relative to next year forecasts.

Figure 3: Weights on current year's density forecast



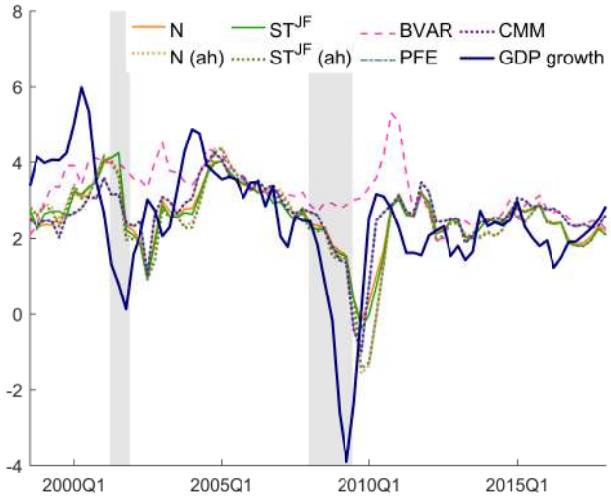
Note: The four panels in the figure depict the estimated combination weights on current year's density forecast corresponding to every quarter for each variable. Qj denotes the jth quarter in the year.

realization, regardless of how the underlying densities are approximated. The BVAR, CMM, as well as the PFE provide unimodal distributions, where the modes for the CMM and PFE are below the actual realization as well as the second, rightmost mode of the mixture distributions. The predictive distribution implied by the BVAR is particularly dispersed.

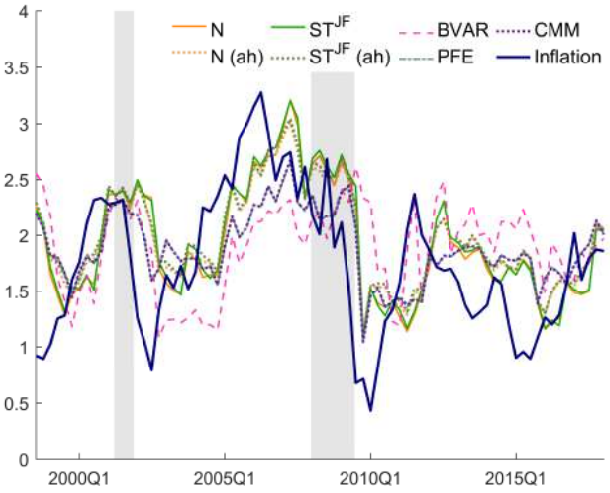
Figure 6 displays the fixed-horizon densities for GDP growth and inflation over time. In Panel a, we can clearly see that during the recent Great Recession, the mean of the predictive distribution of GDP growth decreased and its dispersion increased. Furthermore, the densities are strikingly skewed during that period, in line with the findings of Adrian et al. (2019). In the case of inflation (Panel b), it is remarkable to see that as our estimator "selects" current year's inflation forecast in most periods, the predictive distributions are fairly tight around the actual realizations, yet there is noticeable time-variation in these densities over time.

Figure 4: Mean, standard deviation and skewness of four-quarter-ahead density forecasts

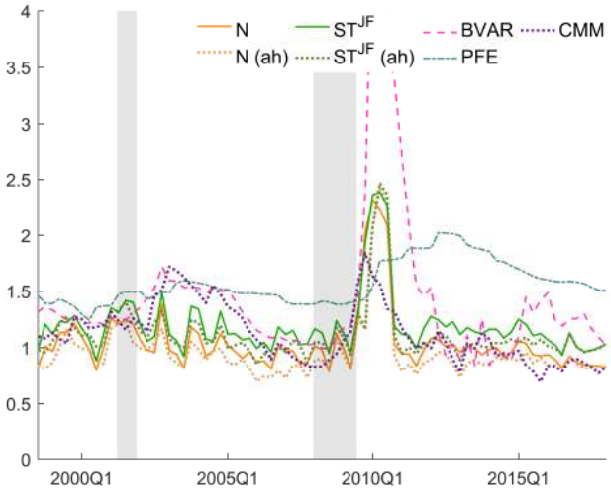
(a) GDP growth, mean



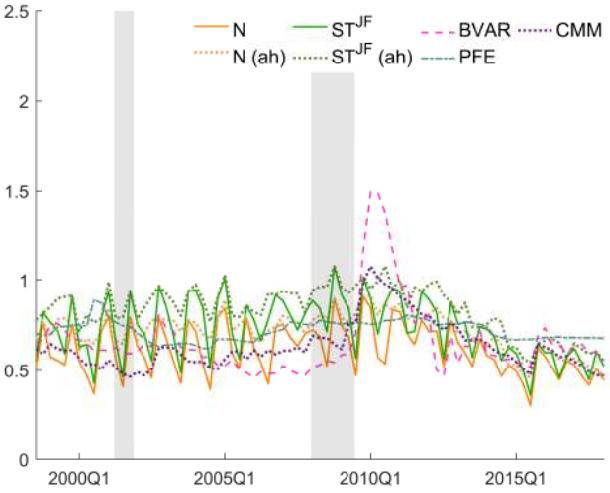
(b) Inflation, mean



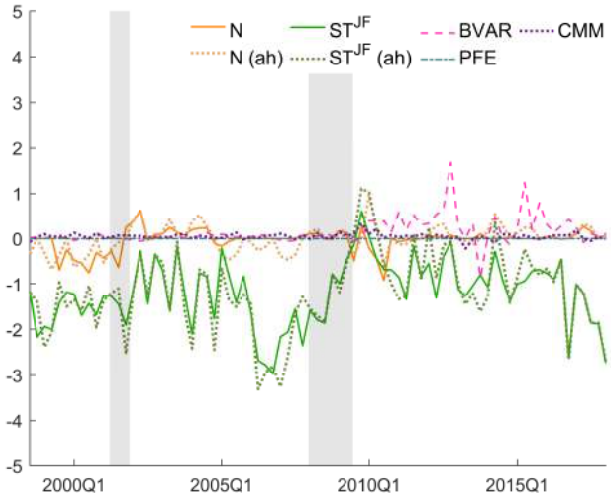
(c) GDP growth, standard deviation



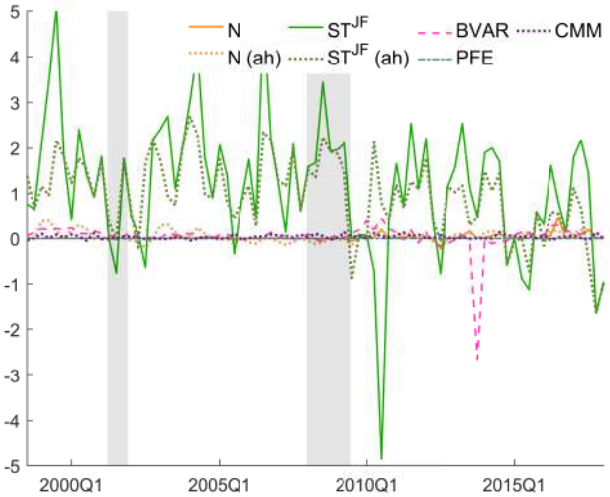
(d) Inflation, standard deviation



(e) GDP growth, skewness

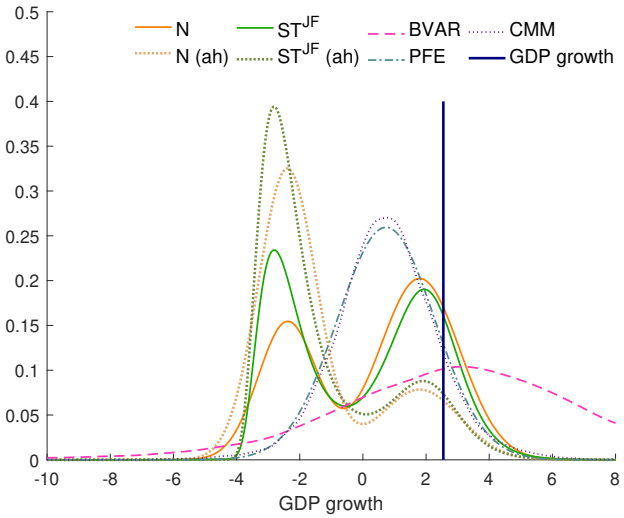


(f) Inflation, skewness



Note: The figures show the mean, standard deviation and skewness (standardized third central moment) of the four-quarter-ahead GDP growth (Panels a, c and e) and inflation (Panels b, d and f) forecasts of various models at the corresponding target dates, ranging from 1998:Q3 to 2018:Q1. For an explanation of the different abbreviations, see the main text. Shaded areas denote NBER recession periods.

Figure 5: Comparison of predictive densities for GDP growth in 2009:Q2



Note: The figure shows the four-quarter-ahead predictive densities of various models for GDP growth, as of 2009:Q2. The solid vertical line indicates the actual realization of GDP growth in 2010:Q1. For an explanation of the different abbreviations, see the main text.

Figure 7 shows various quantiles associated with our combined density for inflation and output growth. For instance, the 90% interval in the figures denotes a two-sided symmetric interval around the median and is defined as the interval between 5% and 95% quantiles. In what follows, we refer to these intervals as equal-tailed ones. In the case of GDP growth, the combinations of normals or skew t distributions are very similar and smooth over time. Additionally, the predictive distributions are fairly tight. On the other hand, inflation forecasts display more high-frequency time-variation, and the combination of skew t distributions are somewhat more dispersed than their normal counterparts.

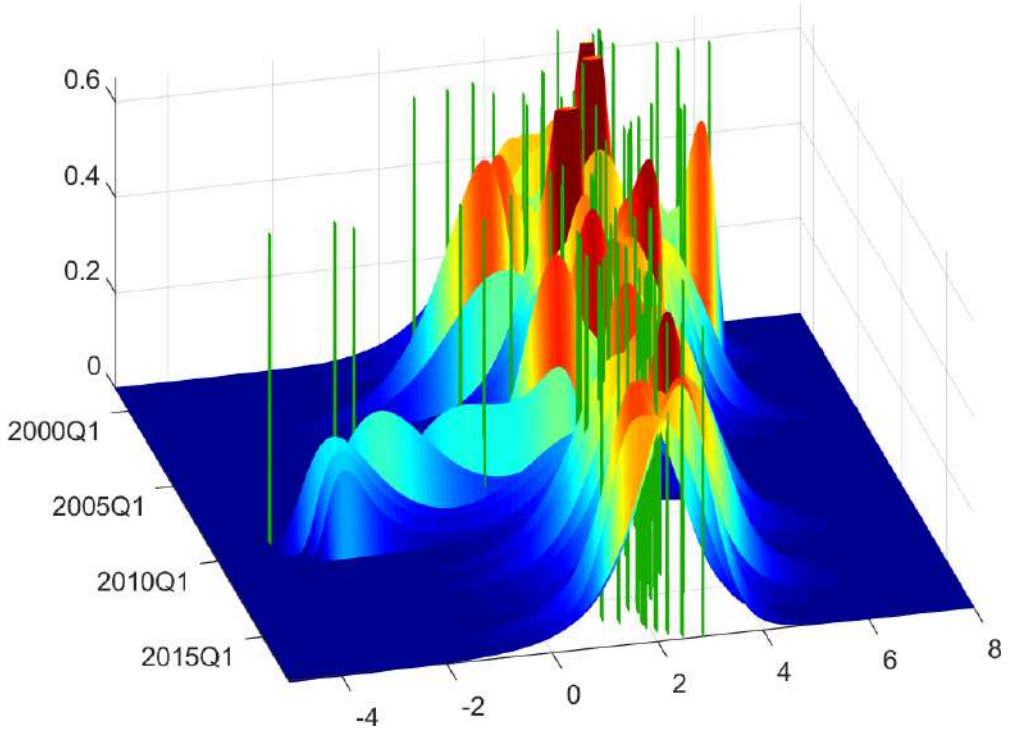
Figure 8 shows the various quantiles of the combined densities using the ad-hoc weights. In the case of inflation, comparing Panels b and d of Figure 8 relative to Panels b and d of Figure 7 reveals that when the densities are combined based on the ad-hoc weights, the quantiles are smoother and the combined density is more dispersed. For GDP growth, however, the density based on ad-hoc weights is much tighter, thus the ad-hoc weights, overall, understate the uncertainty relative to the combined density with optimal weights.

Figure 9 shows that the uncertainty embedded in the BVAR and in the PFE is much higher compared to the combined densities. Furthermore, the PFE predictive distribution of both GDP growth and inflation is relatively stable over time, while the time-varying nature of uncertainty appears to be common to densities obtained with the BVAR. For GDP growth, the CMM model provides somewhat tighter distributions than the skew t distribution with estimated weights (apart from a brief period after the recession in the early 2000s), and it is considerably less dispersed than the PFE model's distribution; the latter highlights the gains from explicitly modeling the time-variation (as in Clark et al., forthcoming) versus simply proxying the variance of the predictive distribution by the mean squared forecast error associated with past forecasts. In the case of inflation, our combination method implies wider distributions (see Panel d in Figure 7) than the CMM model's before the Great Recession, but this reverses after the crisis.[26]

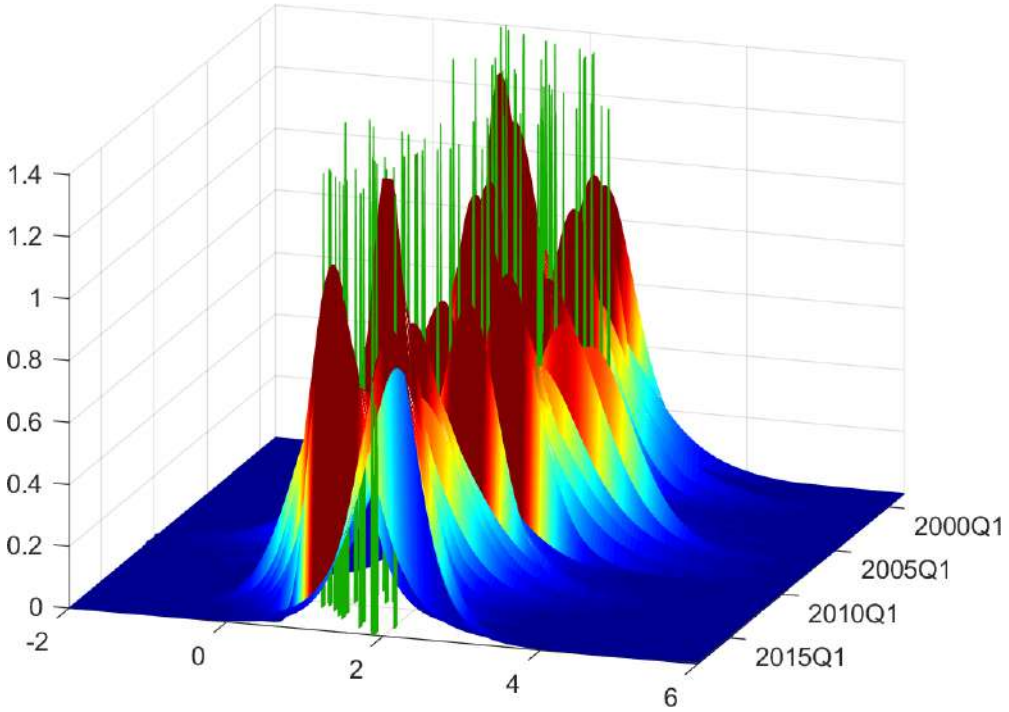
[26] Figure A.6 in Appendix A shows that the results from BVAR and CMM models, estimated recursively, are similar.

Figure 6: Four-quarter-ahead combined skew t predictive densities

(a) GDP growth



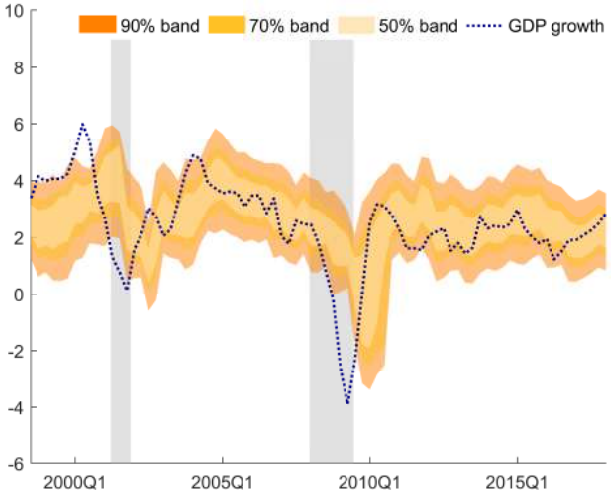
(b) Inflation



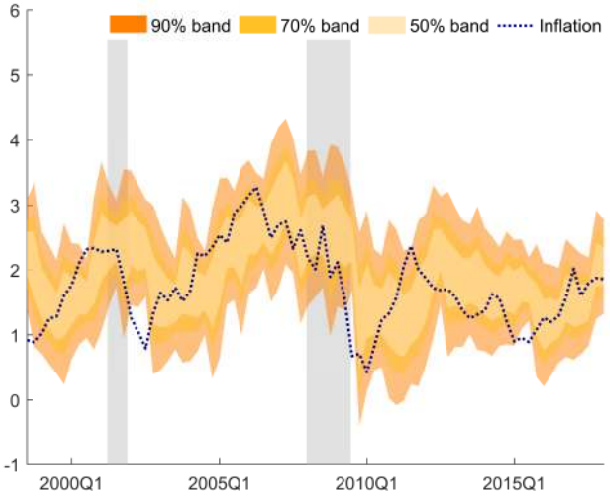
Note: The figures show the combined four-quarter-ahead predictive densities of GDP growth (upper panel) and inflation (lower panel) based on the US SPF, using the proposed weight estimator. The vertical (green) bars mark the realized values of the variable of interest based on the first release. The forecast target dates on the horizontal axis range from 1998:Q3 to 2018:Q1.

Figure 7: Predictive intervals of four-quarter-ahead combined predictive densities, using estimated weights

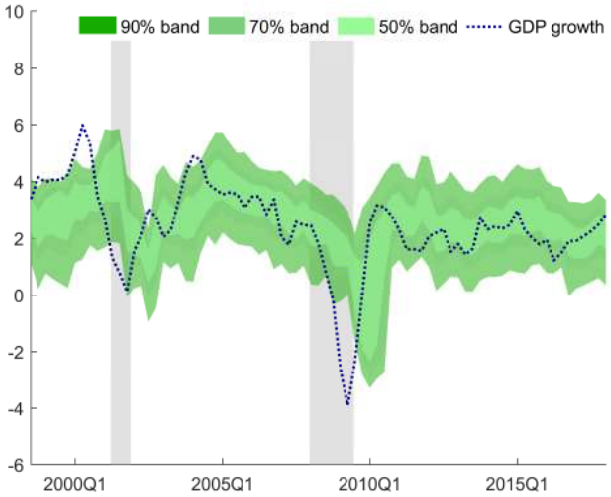
(a) GDP growth, normals



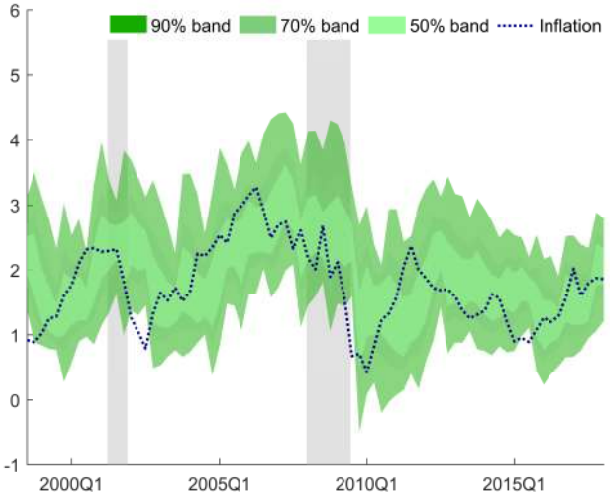
(b) Inflation, normals



(c) GDP growth, skew t



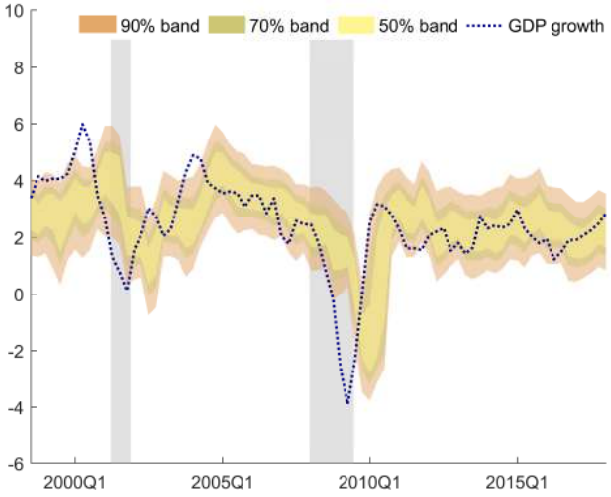
(d) Inflation, skew t



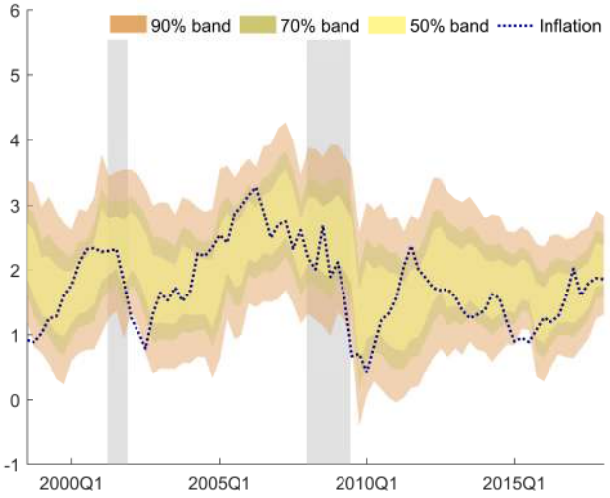
Note: The figure shows 90%, 70%, and 50% bands, corresponding to the 90%, 70%, and 50% equal-tailed predictive intervals of the combined four-quarter-ahead predictive densities for GDP growth (left column) and inflation (right column) based on the US SPF, using the proposed weight estimator. The dotted (blue) lines mark the realized values of the variable of interest according to the first release. The forecast target dates on the horizontal axis range from 1998:Q3 to 2018:Q1. Shaded areas denote NBER recession periods.

Figure 8: Predictive intervals of four-quarter-ahead combined predictive densities, using ad-hoc weights

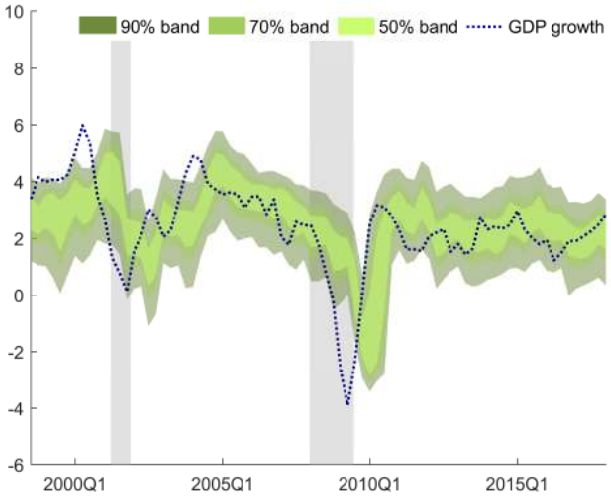
(a) GDP growth, normals



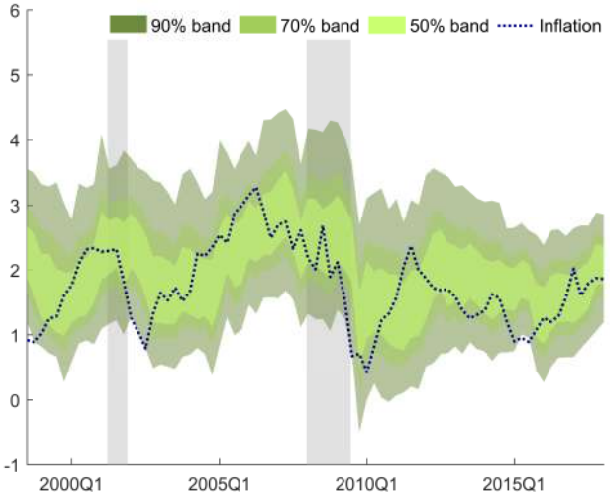
(b) Inflation, normals



(c) GDP growth, skew t



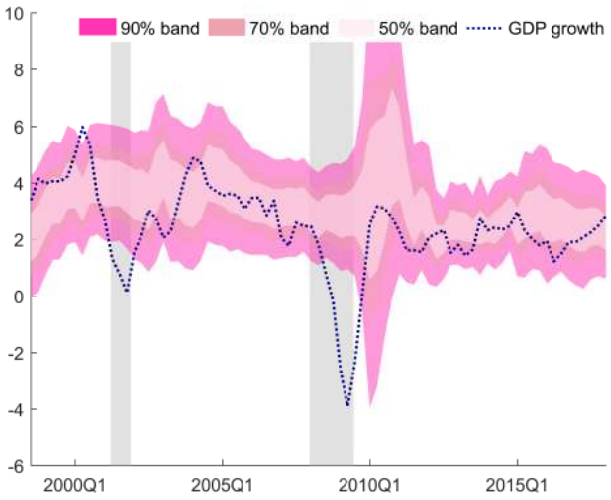
(d) Inflation, skew t



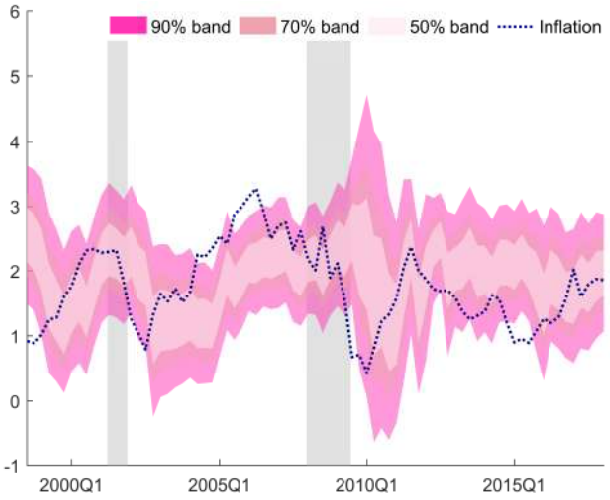
Note: The figure shows 90%, 70%, and 50% bands, corresponding to the 90%, 70%, and 50% equal-tailed predictive intervals of the combined four-quarter-ahead predictive densities for GDP growth (left column) and inflation (right column) based on the US SPF, using ad-hoc weights. The dotted (blue) lines mark the realized values of the variable of interest according to the first release. The forecast target dates on the horizontal axis range from 1998:Q3 to 2018:Q1. Shaded areas denote NBER recession periods.

Figure 9: Predictive intervals of four-quarter-ahead predictive densities of BVAR, PFE and CMM models

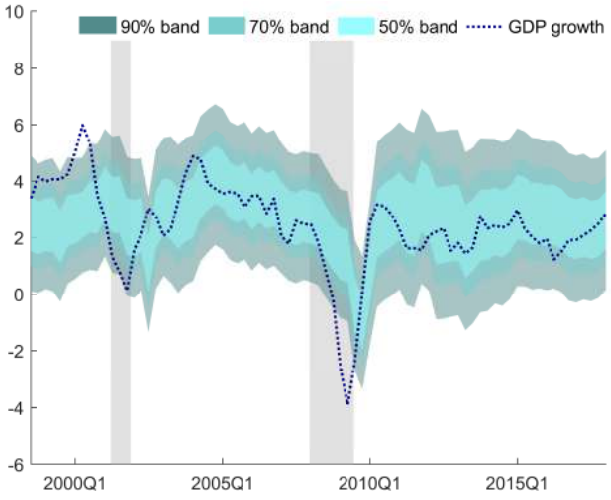
(a) GDP growth, BVAR



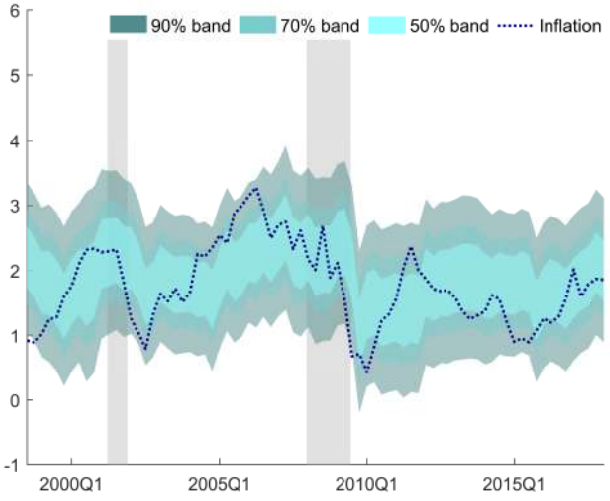
(b) Inflation, BVAR



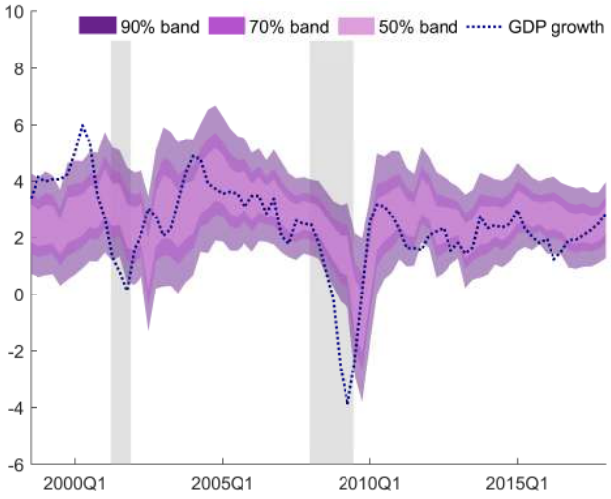
(c) GDP growth, PFE



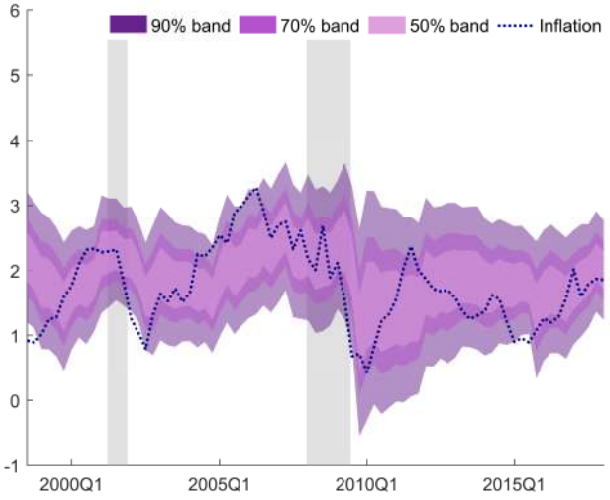
(d) Inflation, PFE



(e) GDP growth, CMM



(f) Inflation, CMM



Note: The figure shows 90%, 70%, 50% bands, corresponding to the Bayesian VAR's (or the PFE model's or the CMM model's) 90%, 70% and 50% equal-tailed predictive intervals for GDP growth and inflation. The dotted blue lines mark the realized values of the variable of interest according to the first release. The forecast target dates on the horizontal axis range from 1998:Q3 to 2018:Q1. Shaded areas denote NBER recession periods.

Figures 7 to 9 communicate uncertainty in terms of quantile-based, equal-tailed forecast intervals. When the predictive density is not unimodal and symmetric, these intervals could mask some information. More particularly, given that in our case the mixture densities (shown in Figure 6) could be skewed and, at times, multi-modal, other summary metrics could potentially be more useful. For instance, as considered in Wallis (1999) and Mitchell and Weale (2019), the Bank of England's fan charts display the highest density regions ("HDR"), referred to as "best critical regions," instead of the equal-tailed prediction intervals. The highest density regions are the intervals of shortest length with a given target coverage, say 90%. When the distributions are unimodal and symmetric, these two measures overlap. To demonstrate how important asymmetry and multi-modality are for the predictive densities we consider, we show the 50%, 70% and 90% highest density regions in Figures 10 to 12.[27] We use the density quantile approach outlined in Hyndman (1996) to calculate these regions.

When we compare Figures 7 and 8 to Figures 10 and 11, respectively, there are some noticeable differences. For instance, in the case of the GDP growth forecasts, right after the Great Recession (at the end of 2009 and in early 2010), the highest density region communicates tighter and, at times, disjoint intervals relative to the equal-tailed intervals with the same coverage. On the other hand, for inflation, the 50% HDR region displays a few disjoint intervals in the first half of the 2000s, including values such as 2% inflation as well as roughly twice as much. As discussed in Wallis (1999), HDR regions would be more informative relative to equal-tailed intervals for an agent with an all-or-nothing loss function (which is typically minimized by the mode of the distribution). When comparing Figure 9 to Figure 12, on the other hand, we find no major differences between the equal-tailed intervals and HDRs.

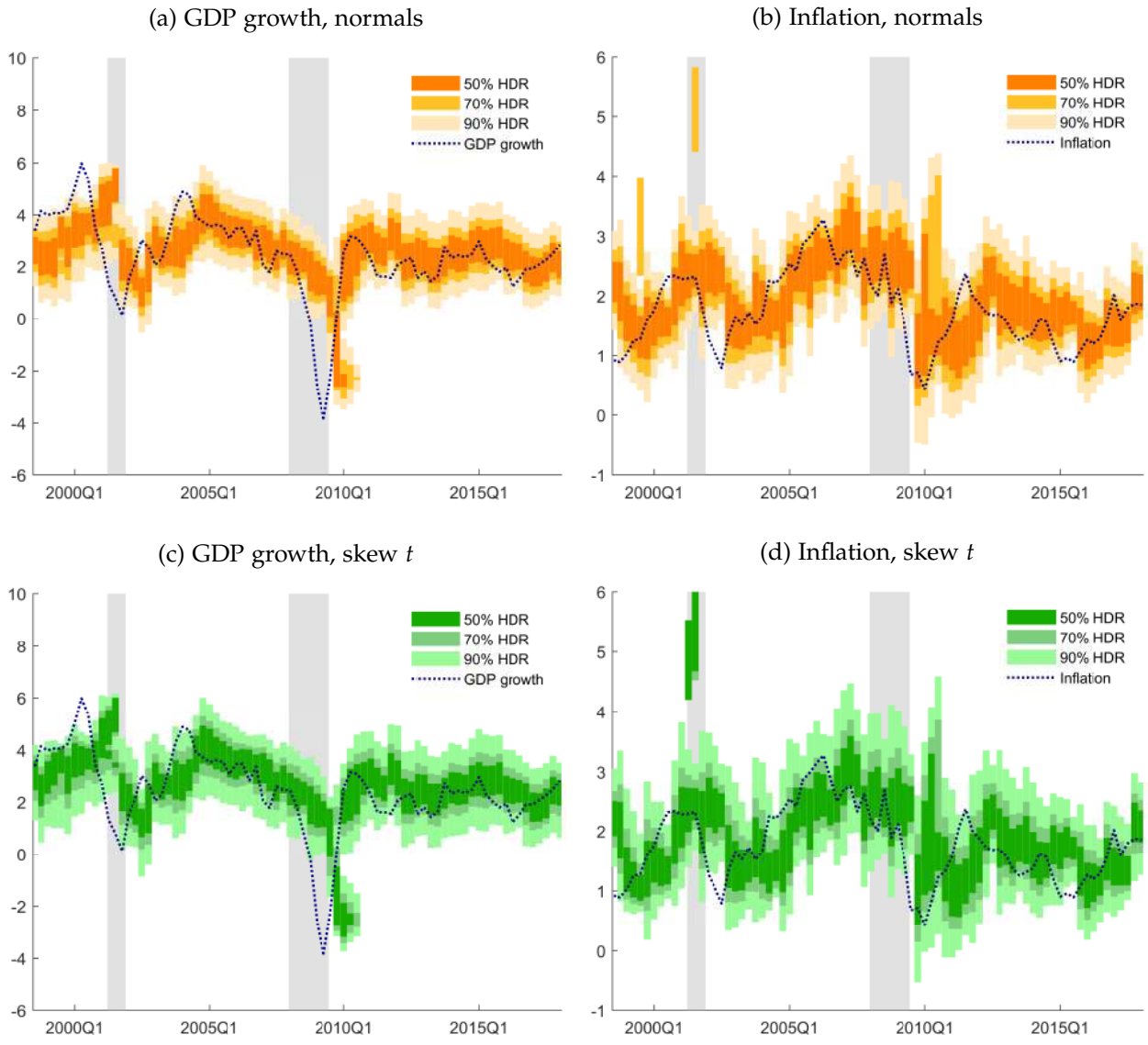
4.1 A formal comparison of fixed-horizon predictive densities

Which fixed-horizon predictive density should researchers use in practice? We compare them by using Rossi and Sekhposyan's (2019) test on the uniformity of the PITs, using both the Kolmogorov–Smirnov (KS) and the Cramér–von Mises (CvM) test statistics with bootstrapped critical values.[28] Table 1 shows the results. In each cell, the test statistic is displayed first, followed by the p-value of testing the null hypothesis of the uniformity of the PIT. The cases in which uniformity cannot be rejected at the 10% level are in bold. We can see that, for GDP growth, uniformity cannot be rejected for the mixtures of both the normal and the skew t distributions when the weights are estimated using our proposed method, while the ad-hoc combination delivers uniform PITs only for GDP growth but not for inflation. Moreover, the BVAR, PFE and CMM models show evidence of incorrect specification for GDP growth according to at least one of the test statistics.

To gain a better understanding of the absolute performance of the various density forecast approaches, we report the empirical coverage rates at the 50% and 70% nominal rates. In practice, for each variable and each forecasting method, we determined the 25th and 75th percentiles (50% nominal rate), and the 15th and 85th percentiles (70% nominal rate) of the predictive distributions in each quarter (as displayed in Figures 7 to 9), and calculated the ratio of cases when the realization of a particular variable fell inside these intervals. Then, we performed a two-sided

[27] By definition, the equal-tailed intervals and the HDRs are the same for the PFE forecasts, hence omitted.
[28] Bootstrapped critical values take into account the serial correlation associated with multi-step-ahead PITs.

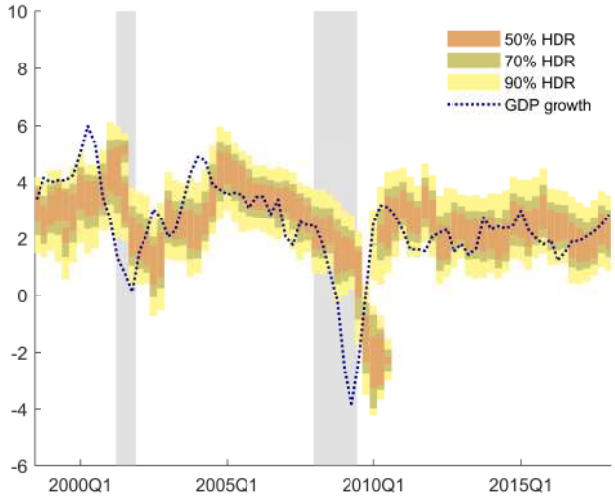
Figure 10: Highest density regions of four-quarter-ahead combined predictive densities, using estimated weights



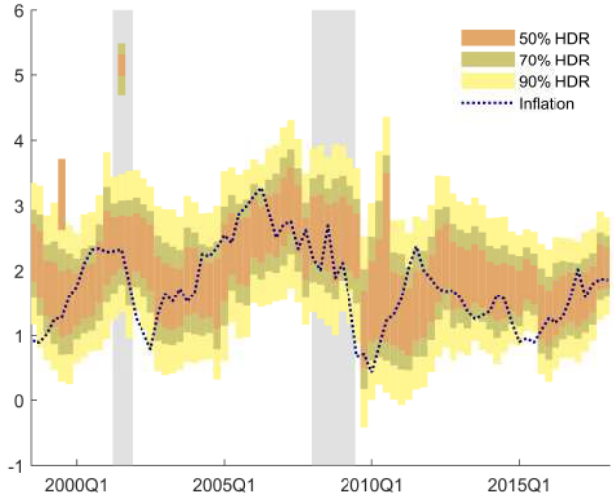
Note: The figure shows 90%, 70%, 50% highest density regions for GDP growth and inflation. The dotted blue lines mark the realized values of the variable of interest. The forecast target dates on the horizontal axis range from 1998:Q3 to 2018:Q1. Shaded areas denote NBER recession periods.

Figure 11: Highest density regions of four-quarter-ahead combined predictive densities, using ad-hoc weights

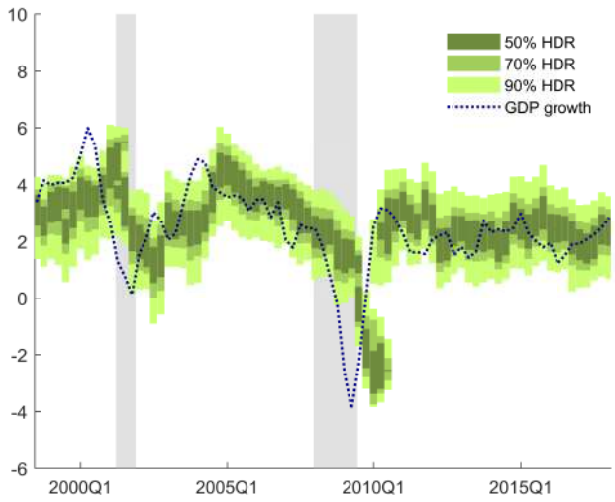
(a) GDP growth, normals



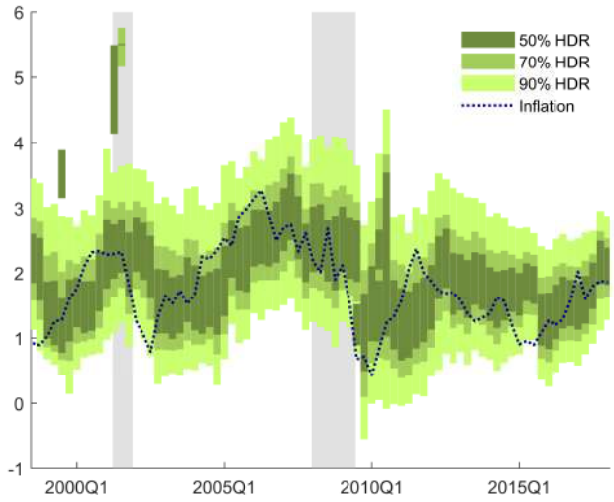
(b) Inflation, normals



(c) GDP growth, skew t

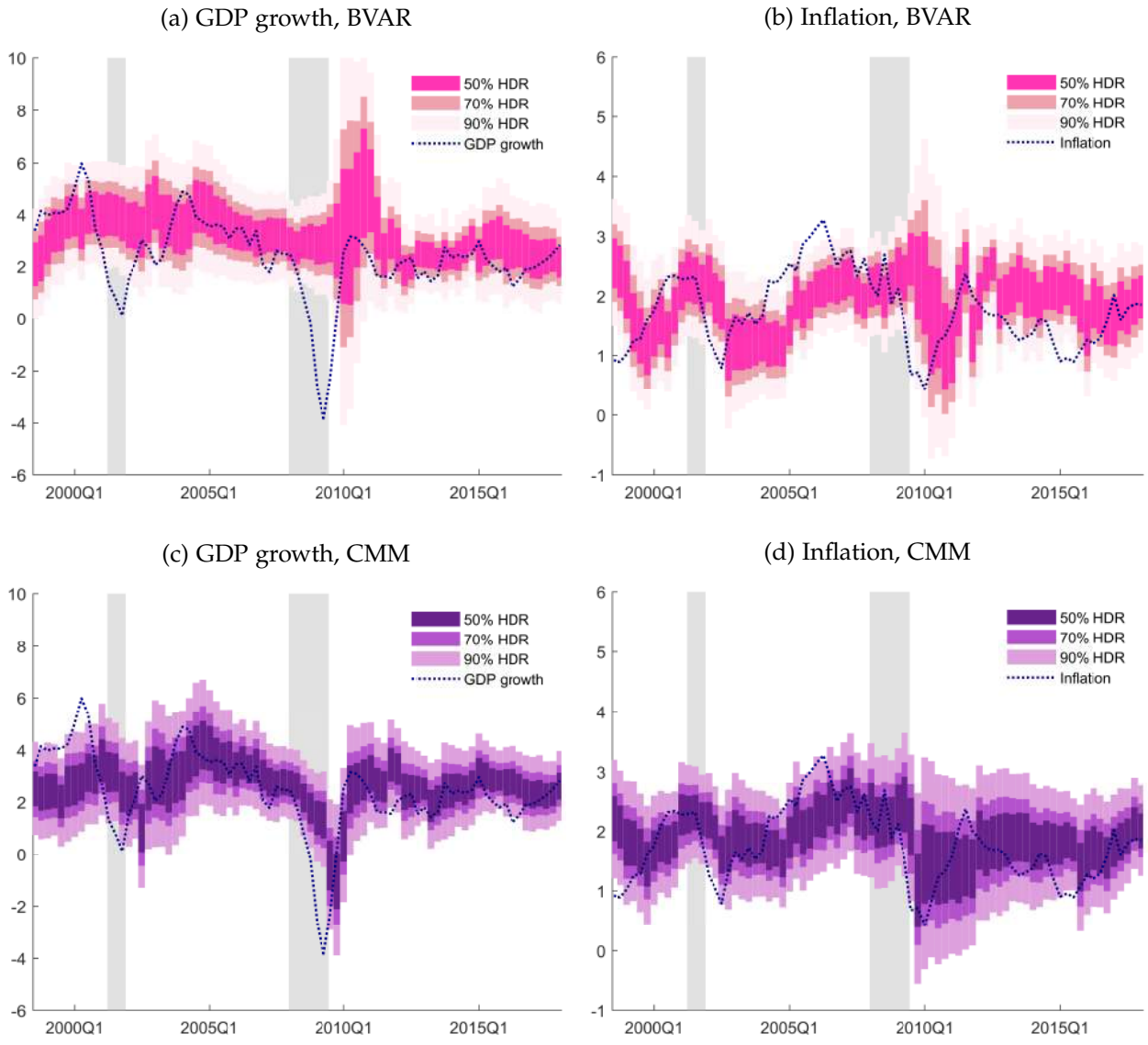


(d) Inflation, skew t



Note: The figure shows 90%, 70%, 50% highest density regions for GDP growth and inflation. The dotted blue lines mark the realized values of the variable of interest. The forecast target dates on the horizontal axis range from 1998:Q3 to 2018:Q1. Shaded areas denote NBER recession periods.

Figure 12: Highest density regions of four-quarter-ahead predictive densities of BVAR and CMM models



Note: The figure shows 90%, 70%, 50% highest density regions for GDP growth and inflation. The dotted blue lines mark the realized values of the variable of interest. The forecast target dates on the horizontal axis range from 1998:Q3 to 2018:Q1. Shaded areas denote NBER recession periods.

Table 1: Absolute forecast evaluation: uniformity of PIT

	GDP growth		Inflation	
	KS	CvM	KS	CvM
N	**0.93**(**0.51**)	**0.19**(**0.60**)	**1.16**(**0.27**)	**0.40**(**0.23**)
ST^{JF}	**0.94**(**0.52**)	**0.25**(**0.48**)	**0.91**(**0.48**)	**0.30**(**0.32**)
N (ah)	**0.96**(**0.47**)	**0.26**(**0.46**)	1.68(0.06)	0.90(0.05)
ST^{JF}(ah)	**1.03**(**0.40**)	**0.29**(**0.42**)	1.71(0.05)	0.82(0.06)
BVAR	2.29(0.00)	1.87(0.00)	**1.27**(**0.28**)	**0.28**(**0.50**)
PFE	1.72(0.05)	**0.62**(**0.12**)	**1.45**(**0.18**)	**0.53**(**0.18**)
CMM	1.72(0.05)	**0.73**(**0.12**)	**1.44**(**0.17**)	**0.46**(**0.25**)

Note: The table displays the Kolmogorov–Smirnov (KS) and Cramér–von Mises (CvM) test statistics and p-values of the null hypothesis of uniformity of PITs (in parentheses) for different target variables (in the column headers) and models (in rows). For an explanation of the different abbreviations, see the main text. The p-values are calculated using the block weighted bootstrap proposed by Rossi and Sekhposyan (2019), with block length $\ell = 4$ and 10,000 bootstrap replications. The cases in which uniformity cannot be rejected at the 10% level are reported in bold. The survey dates range from 1997:Q4 to 2017:Q2, with corresponding realizations between 1998:Q3 and 2018:Q1.

t test to test the null hypothesis that a given coverage rate equals its nominal counterpart. The asymptotic variance is calculated using the Newey and West (1987) HAC estimator with one lag.[29] The results in Table 2 show interesting patterns. For both GDP growth and inflation, the mixture densities with estimated weights deliver correct coverage rates at both the 50% and the 70% nominal levels (at the 10% significance level). However, the ad-hoc mixtures for GDP growth (at 70% nominal coverage), and particularly for inflation (both at 50% and 70% nominal coverage) display incorrect coverage — undercoverage for the former variable and overcoverage for the latter. This is in line with our earlier discussion of Figure 8. For GDP growth, both the BVAR and the CMM model deliver correct coverage, while the PFE model significantly overcovers at the 50% rate. On the other hand, for inflation, only the CMM model, but not the BVAR or the PFE model, displays correct coverage at both rates.

The Continuous Ranked Probability Score (CRPS) has been used in several studies to evaluate competing forecasts (e.g. Clark et al., forthcoming). Formally, for the h-quarter-ahead density forecast made in year t and quarter q using model m, it is defined as

$$\text{CRPS}_{t,q+h}^{(m)} \equiv \int_{-\infty}^{\infty} \left(\widehat{F}_{t,q}^{q+h\,(m)}(y) - \mathbb{1}\left[y_{t,q}^{q+h} \leq y \right] \right)^2 \mathrm{d}y\,, \tag{19}$$

where $\widehat{F}_{t,q}^{q+h\,(m)}(y)$ is the corresponding predictive CDF. The average full-sample CRPS is given by

$$\text{CRPS}^{(m)} \equiv |\mathcal{T}|^{-1} \sum_{t \in \mathcal{T}} \text{CRPS}_{t,q+h}^{(m)}\,. \tag{20}$$

Lower values of the CRPS correspond to better models. For the mixture densities, we numerically calculate the integral in Equation (19), while for the MCMC-based densities, such as those obtained from a BVAR and CMM, we used the empirical CDF-based approximation proposed

[29] The data display low serial correlation.

Table 2: Absolute forecast evaluation: coverage

	GDP growth		Inflation	
	50%	70%	50%	70%
N	**49.4**(**0.92**)	**67.1**(**0.66**)	**55.7**(**0.36**)	**73.4**(**0.54**)
ST^{JF}	**48.1**(**0.77**)	**63.3**(**0.31**)	**53.2**(**0.61**)	**73.4**(**0.52**)
N (ah)	**41.8**(**0.19**)	57.0(0.06)	63.3(0.04)	81.0(0.03)
ST^{JF}(ah)	**41.8**(**0.19**)	57.0(0.06)	60.8(0.09)	81.0(0.03)
BVAR	**46.8**(**0.62**)	**69.6**(**0.95**)	**40.5**(**0.14**)	55.7(0.03)
PFE	65.8(0.02)	**77.2**(**0.22**)	63.3(0.04)	**77.2**(**0.20**)
CMM	**49.4**(**0.93**)	**59.5**(**0.12**)	**51.9**(**0.77**)	**69.6**(**0.95**)

Note: The table displays empirical coverage rates and the two-sided p-values of the null hypothesis that a given coverage rate equals its nominal counterpart (in parentheses) for different target variables at different nominal coverage rates (in the column headers) and models (in rows). For an explanation of the different abbreviations, see the main text. The test statistics were calculated using the Newey and West (1987) HAC estimator with one lag. The cases in which the null hypothesis cannot be rejected at the 10% level are reported in bold. The survey dates range from 1997:Q4 to 2017:Q2, with corresponding realizations between 1998:Q3 and 2018:Q1.

by Krüger et al. (2017). Due to normality, for the PFE model we could analytically calculate the CRPS, following Gneiting and Raftery (2007).

The top panel in Table 3 shows the CRPS of the proposed density combination, along with its competitors'. The bottom panel in Table 3 reports the Diebold and Mariano (1995) test statistics and the corresponding p-values (in parentheses) when equal predictive ability is measured by the CRPS. Negative values mean that the first model is better than the second one. The test statistics were calculated using the Newey and West (1987) HAC variance estimator with one lag (due to low serial correlation). The p-values were calculated based on the standard normal approximation to the asymptotic distribution of the test statistic, with rejection region in the left tail.

As we can see, for GDP growth, our proposed combination scheme achieves the second best CRPS value, after the CMM model. For inflation, the CRPS values are much less dispersed, and the CMM model is the best, very closely followed by the PFE model and the mixture distributions. Furthermore, for GDP growth, our mixture densities are better (although not significantly) than their ad-hoc counterparts, and they significantly outperform the BVAR at the 5% significance level. When predicting inflation, our method beats the BVAR, while the ad-hoc combination and the PFE model are marginally, but not significantly, better than our proposal. The CMM model, while providing the lowest CRPS statistics, is not significantly better than our method. In sum, the mixture densities are sometimes significantly better than their benchmark competitors, and never significantly worse.

4.2 Predicting extreme events

Density forecasts can be used to predict the probability of extreme events, which are of special interests to policymakers. To evaluate how each model performs in forecasting extreme events, we did the following exercise. Using each model, we calculate the probability of two events: GDP growth being lower than (or equal to) 1% and inflation being lower than (or equal to) 1%. The former is an indicator of weak economic activity, while the latter signals "dangerously" low

Table 3: Relative forecast evaluation: CRPS

	GDP growth	Inflation
N	0.75	0.34
ST^{JF}	0.75	0.34
N (ah)	0.79	0.33
ST^{JF}(ah)	0.79	0.33
BVAR	0.90	0.45
PFE	0.76	0.32
CMM	**0.72**	**0.32**
N vs N (ah)	$-0.99(0.16)$	$0.98(0.84)$
ST^{JF} vs ST^{JF}(ah)	$-1.35(0.09)^{*}$	$1.19(0.88)$
N vs BVAR	$-1.93(0.03)^{**}$	$-3.22(0.00)^{***}$
ST^{JF} vs BVAR	$-2.04(0.02)^{**}$	$-3.14(0.00)^{***}$
N vs PFE	$-0.16(0.44)$	$0.74(0.77)$
ST^{JF} vs PFE	$-0.32(0.37)$	$0.94(0.83)$
N vs CMM	$0.84(0.80)$	$0.92(0.82)$
ST^{JF} vs CMM	$0.55(0.71)$	$1.12(0.87)$

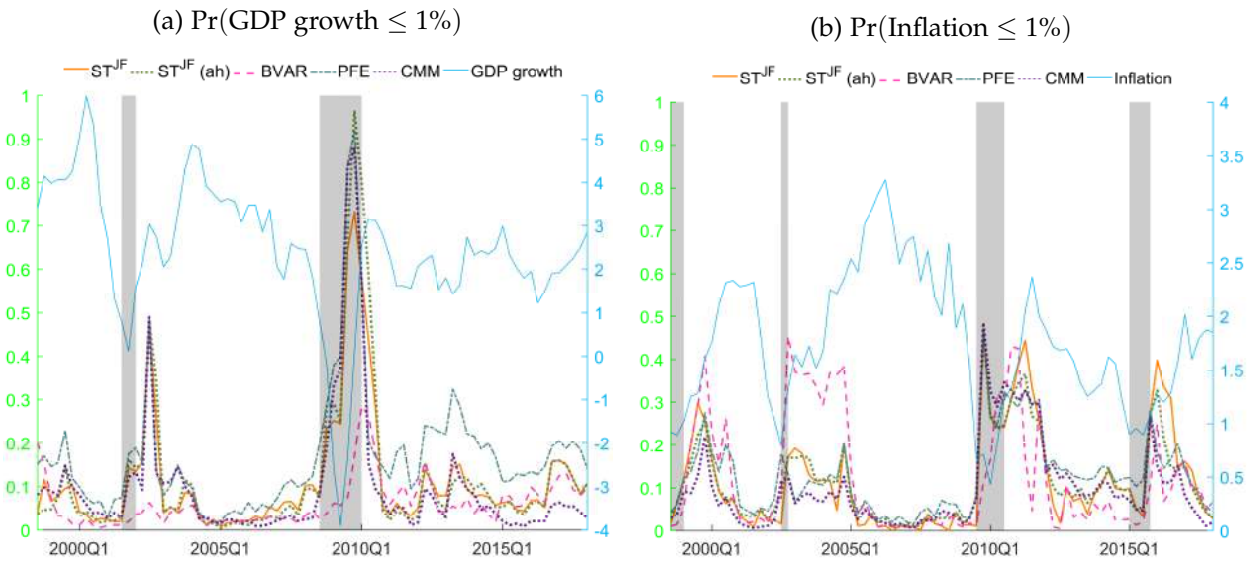
Note: The target variable used for both estimation and forecast evaluation is shown in the column headers. The top panel displays the Continuous Ranked Probability Score (CRPS) of various density combination methods in the rows. For each variable, the lowest value is in bold. For an explanation of the different abbreviations, please see the main text. The bottom panel displays the Diebold and Mariano (1995) test statistics and p-values (in parentheses, with rejection region in the left tail) comparing predictive accuracy measured by the CRPS. Negative values indicate that the first method outperforms the second one, *, ** and *** denote rejection at the 10%, 5% and 1% significance level, respectively. The test statistics were calculated using the Newey and West (1987) HAC estimator with one lag. The survey dates range from 1997:Q4 to 2017:Q2, with corresponding realizations between 1998:Q3 and 2018:Q1.

inflation. Figure 13 shows the probabilities for each variable (scaling on the left axis), along with the actual realizations of the variable (scaling on the right axis).[30] The shaded grey areas highlight the periods when the predicted event did in fact occur. For GDP growth, Panel a demonstrates that the PFE model consistently signaled a relatively high probability, in line with their overly dispersed predictive distributions, while the density combination models displayed a considerably lower "baseline" probability in tranquil times. Interestingly, the CMM model's implied probability moves very closely together with that of the mixture using estimated weights. Furthermore, all models react with a lag – and the BVAR did not detect the transitory economic downturn in the early 2000s. When forecasting low inflation instead, we can see that the spikes in Panel b in Figure 13 (and especially the BVAR model's predictions) actually correspond to episodes of low inflation, although the BVAR's predictions show considerable persistence. On the other hand, the density combinations' and even the PFE model's forecasts adapt fairly quickly both before and after low inflation periods. We can see that the CMM model's predictions are very similar to those of the combination methods.

We formally evaluated each model's predictions for the aforementioned extreme events using the Brier (or quadratic) score (Gneiting and Raftery, 2007). For the h-quarter-ahead density forecast made in year t and quarter q using model m and at threshold k (in our case, $k = 1\%$), it is defined as

[30] The probabilities implied by the mixtures of normal distributions are available upon request.

Figure 13: Predicted probabilities of low growth and low inflation



Note: The figure shows according to each model the probabilities of either GDP growth or inflation being less than or equal to 1% (left axis), along with the actual realization of the respective variable (solid blue line, right axis). For an explanation of the different abbreviations, see the main text. The forecast target dates on the horizontal axis range from 1998:Q3 to 2018:Q1. Shaded grey areas denote the periods when the predicted event (e.g. GDP growth ≤ 1%) did in fact occur.

$$\mathrm{BS}^{(m)}_{t,q+h}(k) \equiv \left(\widehat{F}^{q+h\,(m)}_{t,q}(k) - \mathbb{1}\left[y^{q+h}_{t,q} \leq k\right]\right)^2, \tag{21}$$

which is precisely the integrand in Equation (19). Lower values of the Brier score correspond to better predictions. The full-sample Brier score is defined analogously as

$$\mathrm{BS}^{(m)}(k) \equiv |\mathcal{T}|^{-1} \sum_{t\in\mathcal{T}} \mathrm{BS}^{(m)}_{t,q+h}(k)\,. \tag{22}$$

In the upper panel of Table 4 we can see the Brier scores of each model for predicting economic downturns and low inflation. For both events, the CMM model is the most precise (in bold), closely followed by the model based on past forecast errors. The lower panel of Table 4 displays the Diebold and Mariano (1995) test statistics for equal predictive ability based on the Brier score and the corresponding p-values (in parentheses). The p-values are calculated analogously to the forecast comparison based on the CRPS earlier. When forecasting GDP growth, the combination schemes with estimated weights outperform the BVAR, significantly so when combining normals, while the rest of the comparisons are not significant (although the CMM model would be significantly better than the combinations with estimated weights, based on a two-sided test). When predicting low inflation, we can see again that most of the differences are not significant at the 5% level (although the ad-hoc weighting scheme, the PFE and the CMM models would outperform our proposed method). On the other hand, our proposed weight estimation scheme significantly outperforms the BVAR. However, these results should be interpreted with caution, as Figure 13 shows that these were indeed very rare events. We expect that the predictability of tail events to improve if weights are selected to obtain correct calibration in the tails.

Table 4: Relative forecast evaluation: Brier score

	GDP growth $\leq 1\%$	Inflation $\leq 1\%$
N	0.072	0.124
ST^{JF}	0.073	0.122
N (ah)	0.076	0.115
ST^{JF}(ah)	0.076	0.113
BVAR	0.096	0.140
PFE	0.070	0.108
CMM	**0.063**	**0.108**
N vs N (ah)	$-0.59(0.28)$	$2.63(1.00)$
ST^{JF} vs ST^{JF}(ah)	$-0.55(0.29)$	$2.70(1.00)$
N vs BVAR	$-1.33(0.09)^{*}$	$-1.62(0.05)^{*}$
ST^{JF} vs BVAR	$-1.26(0.10)$	$-1.78(0.04)^{**}$
N vs PFE	$0.21(0.58)$	$2.37(0.99)$
ST^{JF} vs PFE	$0.36(0.64)$	$2.30(0.99)$
N vs CMM	$2.13(0.98)$	$3.26(1.00)$
ST^{JF} vs CMM	$2.14(0.98)$	$2.92(1.00)$

Note: The target variable used for both estimation and forecast evaluation and the corresponding extreme event are shown in the column headers. The top panel displays the Brier score of various density combination methods in the rows. For each variable, the lowest value is in bold. For an explanation of the different abbreviations, please see the main text. The bottom panel displays the Diebold and Mariano (1995) test statistics and p-values (in parentheses, with rejection region in the left tail) comparing predictive accuracy measured by the Brier score. Negative values indicate that the first method outperforms the second one, *, ** and *** denote rejection at the 10%, 5% and 1% significance level, respectively. The test statistics were calculated using the Newey and West (1987) HAC estimator with one lag. The survey dates range from 1997:Q4 to 2017:Q2, with corresponding realizations between 1998:Q3 and 2018:Q1.

5 Conclusion

This paper proposes a methodology to construct fixed-horizon density forecasts by combining fixed-event ones. Survey density forecasts are an important application for this methodology; in particular the US Survey of Professional Forecasters, for which fixed-horizon predictive densities are not available. We show that, by estimating the weights according to our criterion, the fixed-horizon combined predictive density appears to be correctly calibrated out-of-sample. In relative terms, our combination scheme is fairly competitive and on par with the historical distribution of point forecast errors or a stochastic volatility model fitted to forecast errors, and outperforms a small Bayesian VAR with stochastic volatility. The improved performance is more pronounced for GDP growth — but, when measured against the BVAR, also for inflation. Hence, our approach makes the SPF densities more useful for policy analysis and communication.

References

Adrian, T., Boyarchenko, N., and Giannone, D. (2019). Vulnerable Growth. *American Economic Review*, 109(4):1263–1289.

Andreou, E., Ghysels, E., and Kourtellos, A. (2010). Regression models with mixed sampling frequencies. *Journal of Econometrics*, 158(2):246–261.

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

Azzalini, A. and Capitanio, A. (2003). Distributions Generated by Perturbation of Symmetry with Emphasis on a Multivariate Skew t-Distribution. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(2):367–389.

Azzalini, A. and Capitanio, A. (2014). *The Skew-Normal and Related Families*. Number 3 in Institute of Mathematical Statistics Monographs. Cambridge University Press, Cambridge. OCLC: 882941451.

Bai, J. (2003). Testing Parametric Conditional Distributions of Dynamic Models. *Review of Economics and Statistics*, 85(3):531–549.

Clark, T. E., McCracken, M., and Mertens, E. (forthcoming). Modeling Time-Varying Uncertainty of Multiple-Horizon Forecast Errors. *Review of Economics and Statistics*.

Clark, T. E. and Ravazzolo, F. (2015). Macroeconomic Forecasting Performance under Alternative Specifications of Time-Varying Volatility. *Journal of Applied Econometrics*, 30(4):551–575.

Clements, M. P. (2004). Evaluating the Bank of England Density Forecasts of Inflation. *The Economic Journal*, 114(498):844–866.

Clements, M. P. (2014a). Forecast Uncertainty—Ex Ante and Ex Post: U.S. Inflation and Output Growth. *Journal of Business & Economic Statistics*, 32(2):206–216.

Clements, M. P. (2014b). Probability distributions or point predictions? Survey forecasts of US output growth and inflation. *International Journal of Forecasting*, 30(1):99–117.

Clements, M. P. (2018). Are macroeconomic density forecasts informative? *International Journal of Forecasting*, 34(2):181–198.

Corradi, V. and Swanson, N. R. (2006). Chapter 5 Predictive Density Evaluation. In Elliott, G., Granger, C. W. J., and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 1, pages 197–284. Elsevier.

Czado, C., Gneiting, T., and Held, L. (2009). Predictive Model Assessment for Count Data. *Biometrics*, 65(4):1254–1261.

D'Amico, S. and Orphanides, A. (2008). Uncertainty and disagreement in economic forecasting. Finance and Economics Discussion Series 2008-56, Washington: Board of Governors of the Federal Reserve System.

Del Negro, M., Casarin, R., and Bassetti, F. (2018). A Bayesian Approach for Inference on Probabilistic Surveys. Working Paper.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review*, 39(4):863–883.

Diebold, F. X. and Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3):253.

Dovern, J., Fritsche, U., and Slacalek, J. (2012). Disagreement among forecasters in G7 countries. *Review of Economics and Statistics*, 94(4):1081–1096.

Engelberg, J., Manski, C. F., and Williams, J. (2009). Comparing the Point Predictions and Subjective Probability Distributions of Professional Forecasters. *Journal of Business & Economic Statistics*, 27(1):30–41.

Federal Reserve Bank of Philadelphia (2017). Documentation of the Survey of Professional Forecasters. Technical report.

Galbraith, J. W. and van Norden, S. (2012). Assessing gross domestic product and inflation probability forecasts derived from Bank of England fan charts. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(3):713–727.

Ganics, G. (2017). Optimal density forecast combinations. Working Paper No. 1751, Banco de España.

Ghysels, E., Sinko, A., and Valkanov, R. (2007). MIDAS Regressions: Further Results and New Directions. *Econometric Reviews*, 26(1):53–90.

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

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Granger, C. W. J. and Pesaran, M. H. (2000). A decision theoretic approach to forecast evaluation. In Chan, W.-S., Li, W. K., and Tong, H., editors, *Statistics and Finance: An Interface*, Proceedings of the Hong Kong International Workshop on Statistics in Finance, pages 261–278. Imperial College Press.

Hyndman, R. J. (1996). Computing and Graphing Highest Density Regions. *The American Statistician*, 50(2):120–126.

Jones, M. C. and Faddy, M. J. (2003). A skew extension of the t-distribution, with applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):159–174.

Kheifets, I. and Velasco, C. (2017). New goodness-of-fit diagnostics for conditional discrete response models. *Journal of Econometrics*, 200(1):135–149.

Knüppel, M. and Vladu, A. L. (2016). Approximating fixed-horizon forecasts using fixed-event forecasts. Discussion Paper No. 28, Deutsche Bundesbank.

Krüger, F., Lerch, S., Thorarinsdottir, T., and Gneiting, T. (2017). Probabilistic Forecasting and Comparative Model Assessment Based on Markov Chain Monte Carlo Output. *ArXiv e-prints*.

Manzan, S. (2015). Forecasting the Distribution of Economic Variables in a Data-Rich Environment. *Journal of Business & Economic Statistics*, 33(1):144–164.

Manzan, S. (2017). Are Professional Forecasters Bayesian? Working Paper.

Mitchell, J. and Weale, M. (2019). Forecasting with Unknown Unknowns: Censoring and Fat Tails on the Bank of England's Monetary Policy Committee. Technical Report 27, Economic Modelling and Forecasting Group.

Newey, W. K. and West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3):703–708.

Patton, A. J. and Timmermann, A. (2010). Why do forecasters disagree? Lessons from the term structure of cross-sectional dispersion. *Journal of Monetary Economics*, 57(7):803–820.

Pettenuzzo, D., Timmermann, A., and Valkanov, R. (2016). A MIDAS approach to modeling first and second moment dynamics. *Journal of Econometrics*, 193(2):315–334.

Reifschneider, D. and Tulip, P. (2017). Gauging the Uncertainty of the Economic Outlook Using Historical Forecasting Errors: The Federal Reserve's Approach. Finance and Economics Discussion Series No. 2017-020, Washington: Board of Governors of the Federal Reserve System.

Rosenblatt, M. (1952). Remarks on a Multivariate Transformation. *Ann. Math. Statist.*, 23(3):470–472.

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

Rossi, B. and Sekhposyan, T. (2014). Evaluating predictive densities of US output growth and inflation in a large macroeconomic data set. *International Journal of Forecasting*, 30(3):662–682.

Rossi, B. and Sekhposyan, T. (2019). Alternative tests for correct specification of conditional predictive densities. *Journal of Econometrics*, 208(2):638–657.

Rossi, B., Sekhposyan, T., and Soupre, M. (2017). Understanding the Sources of Macroeconomic Uncertainty. Manuscript.

Wallis, K. F. (1999). Asymmetric density forecasts of inflation and the Bank of England's fan chart. *National Institute Economic Review*, 167(1):106–112.

Zarnowitz, V. and Lambros, L. A. (1987). Consensus and Uncertainty in Economic Prediction. *Journal of Political Economy*, 95(3):591–621.

Appendix

Appendix A Robustness Checks

A.1 An alternative skew t distribution, and recursively estimated BVAR and CMM models

As explained in Section 3, the probabilistic GDP growth and inflation forecasts in the US SPF are recorded in the form of probabilities assigned to pre-specified bins. However, for our analysis it is necessary to have continuous predictive distributions as mixture components. As we mentioned, several distributions are used in the literature, with the normal being the simplest and most popular choice. Given the importance of skewed predictive distributions, based on both a number of visibly skewed SPF histograms and the recent paper by Adrian et al. (2019), we also performed our analysis using Jones and Faddy's (2003) skew t distribution. However, this is not the only skewed variant of distributions related to Student's t distribution. In the statistics literature, the skew t distribution proposed by Azzalini and Capitanio (2003) is a popular choice. Therefore, we examined the robustness of our results by performing the analysis in Section 3 using Azzalini and Capitanio's (2003) skew t distribution. In its general form with location parameter μ, scale parameter $\sigma > 0$, skewness parameter α and degrees of freedom parameter $\nu > 0$, its PDF at $x \in \mathbb{R}$ is given by

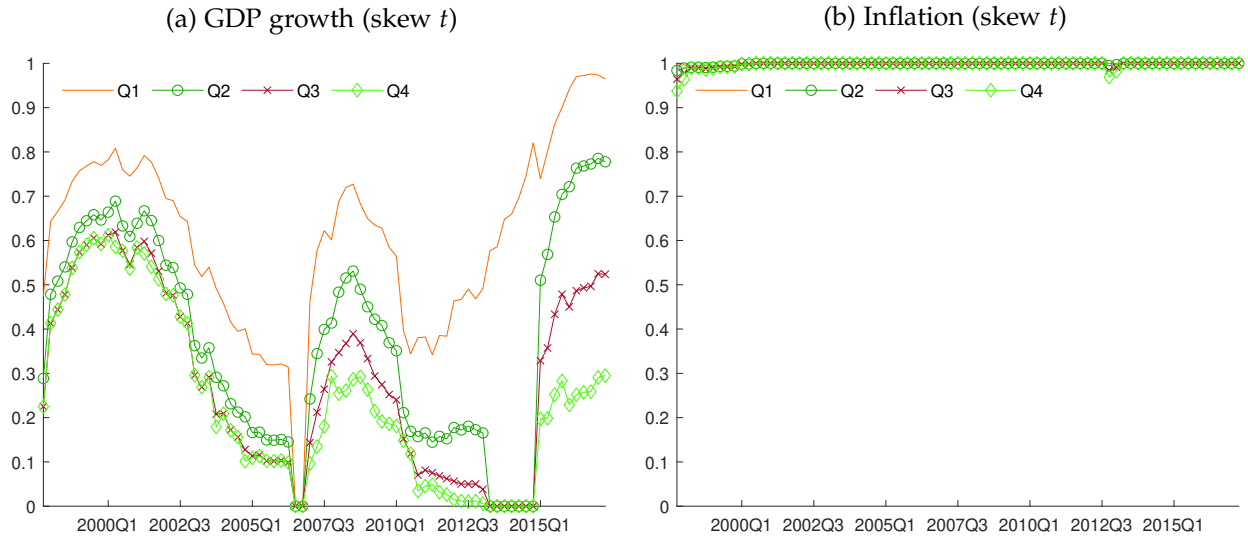
$$f(x;\mu,\sigma,\alpha,\nu) = \frac{2}{\sigma} t_\nu\left(\frac{x-\mu}{\sigma}\right) T_{\nu+1}\left(\alpha\frac{x-\mu}{\sigma}\sqrt{\frac{\nu+1}{\nu+\left(\frac{x-\mu}{\sigma}\right)^2}}\right), \tag{23}$$

where $t_\nu(\cdot)$ and $T_{\nu+1}(\cdot)$ are the PDF of Student's t distribution with degrees of freedom parameter ν, and the CDF of Student's t distribution with degrees of freedom parameter $\nu+1$, respectively.

Let $\theta = (\mu,\sigma,\alpha,\nu)'$ collect the parameters of this distribution. Unfortunately, the CDF of this skew t distribution cannot be expressed in such a simple form as the CDF of Jones and Faddy's (2003) skew t distribution. Therefore, we numerically calculated the integral of the PDF when fitting the CDF of Azzalini and Capitanio's (2003) skew t distribution to the empirical CDFs of the SPF predictions. Furthermore, we restricted the degrees of freedom parameter ν to be greater than or equal to 4 to ensure the existence and finiteness of the fourth moment of the fitted distribution. Hence, we considered the parameter space $\Theta_{\text{AC}} = \mathbb{R}\times\mathbb{R}^+\times\mathbb{R}\times(4,\infty)$. In the following analysis, we used the abbreviation ST^{AC} to index models whose mixture components are skew t distributions of this form.

First, the estimated weights associated with the Azzalini and Capitanio (2003) skew t distribution shown in Figure A.1 are very similar to their counterparts in Figure 3.

Figure A.1: Weights on current year's density forecast with Azzalini and Capitanio's (2003) skew t distribution



Note: The two panels in the figure depict the estimated combination weights on current year's density forecast corresponding to every quarter for each variable. Qj denotes the jth quarter in the year.

Figure A.2 shows that when using Azzalini and Capitanio's (2003) skew t distribution, the means and standard deviations of the mixture distribution are very similar to the case when using Jones and Faddy's (2003) skew t distribution in the main text, both for GDP growth and inflation.

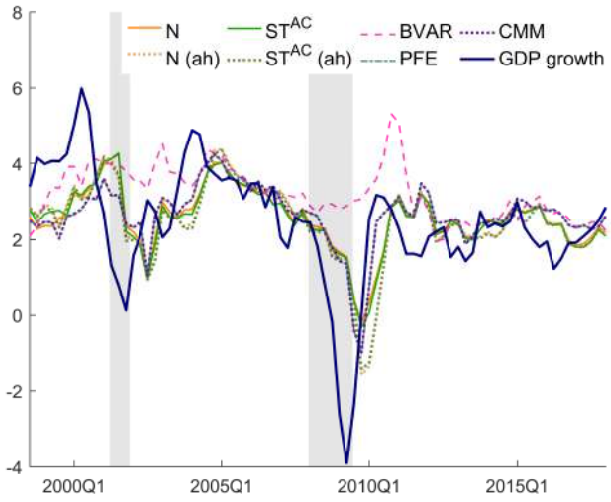
Figures A.3 and A.4 show the predictive intervals obtained as mixtures of Azzalini and Capitanio's (2003) skew t distribution, for both GDP growth and inflation. As we can see, the predictive intervals are visually indistinguishable from the ones in the main text obtained using mixtures of Jones and Faddy's (2003) skew t distribution. The same is true about the predicted probabilities in Figure A.5.

Figure A.6 shows the predictive bands of BVAR and CMM models, where the parameters are estimated in a recursive fashion, where the first estimation window coincides with the first rolling window.

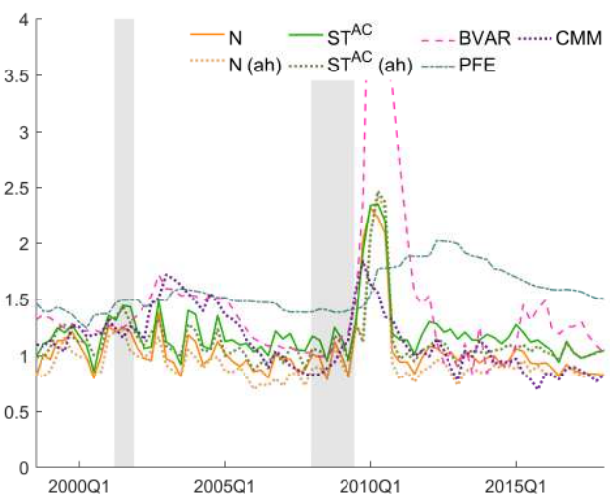
Tables A.1 to A.4 display the same forecast evaluation statistics as in the main text, adding the results with the Azzalini and Capitanio (2003) distribution. Furthermore, we show results for the BVAR and CMM models, where the parameters are estimated based on a recursive estimation scheme. As we can see, the main conclusions are unchanged.

Figure A.2: Mean and standard deviation of four-quarter-ahead GDP growth and inflation forecasts

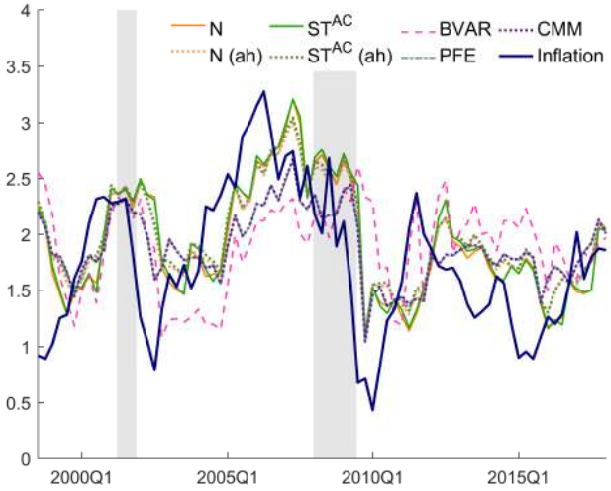
(a) GDP growth, mean



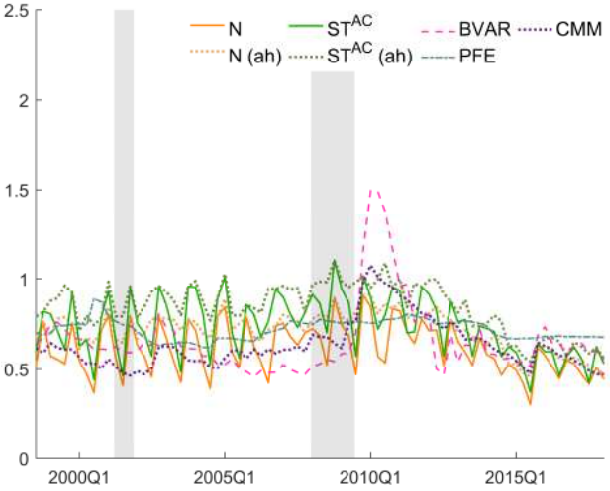
(b) GDP growth, standard deviation



(c) Inflation, mean



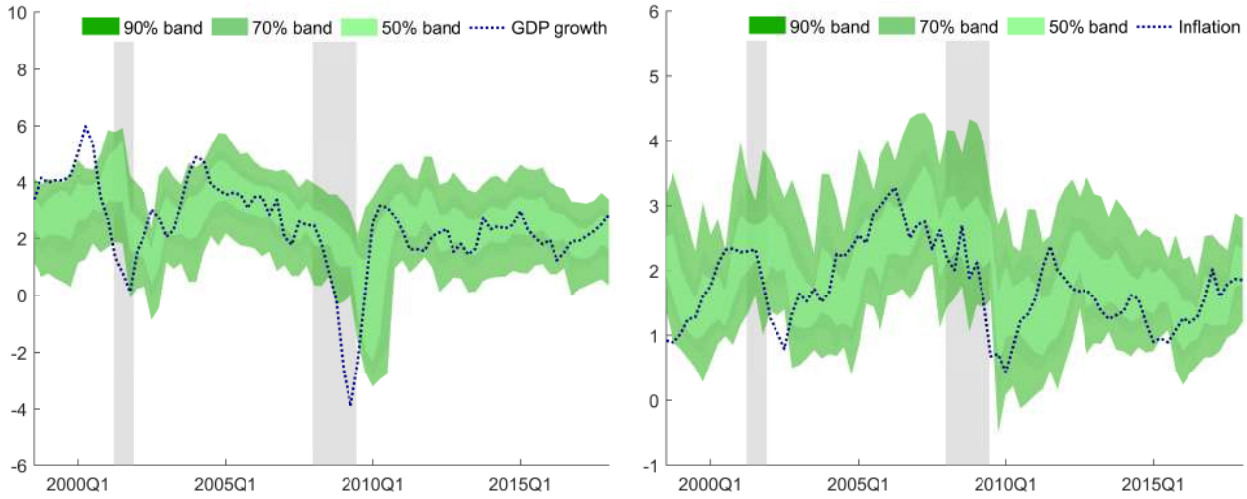
(d) Inflation, standard deviation



Note: The figures show the mean and the standard deviation of the four-quarter-ahead GDP growth forecasts (subfigures a and b) and inflation forecasts (subfigures c and d) of various methods at the corresponding *target* dates ranging from 1998:Q3 to 2018:Q1. For an explanation of the different abbreviations, please see the main text. Shaded areas are NBER recession periods.

Figure A.3: Predictive intervals of four-quarter-ahead combined predictive densities, mixtures of Azzalini and Capitanio's (2003) skew t distribution using estimated weights

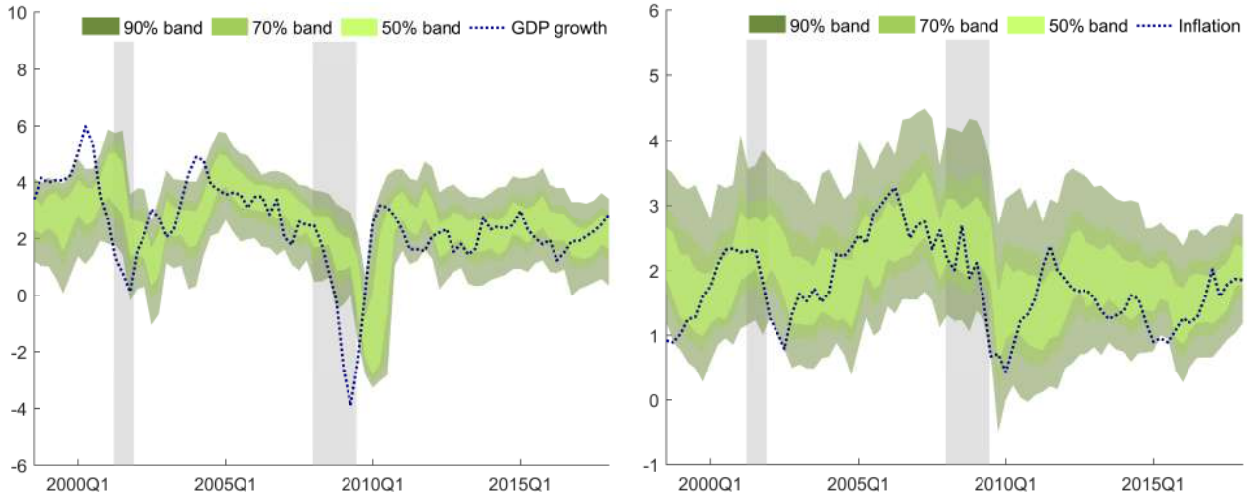
(a) GDP growth, skew t (b) Inflation, skew t



Note: The figure shows 90%, 70%, and 50% bands, corresponding to the 90%, 70%, and 50% equal-tailed predictive intervals of the combined four-quarter-ahead predictive densities for GDP growth (left column) and inflation (right column) based on the US SPF, using the proposed weight estimator. The dotted (blue) lines mark the realized values of the variable of interest according to the first release. The forecast target dates on the horizontal axis range from 1998:Q3 to 2018:Q1. Shaded areas denote NBER recession periods.

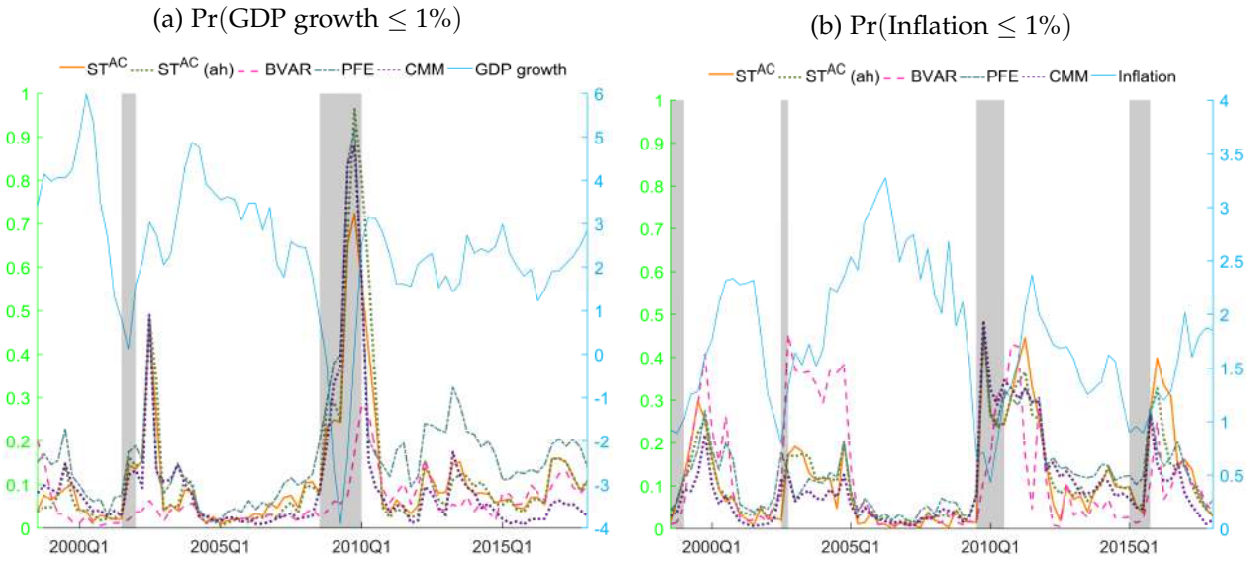
Figure A.4: Predictive intervals of four-quarter-ahead combined predictive densities, mixtures of Azzalini and Capitanio's (2003) skew t distribution using ad-hoc weights

(a) GDP growth, skew t (b) Inflation, skew t



Note: The figure shows 90%, 70%, and 50% bands, corresponding to the 90%, 70%, and 50% equal-tailed predictive intervals of the combined four-quarter-ahead predictive densities for GDP growth (left column) and inflation (right column) based on the US SPF, using ad-hoc weights. The dotted (blue) lines mark the realized values of the variable of interest according to the first release. The forecast target dates on the horizontal axis range from 1998:Q3 to 2018:Q1. Shaded areas denote NBER recession periods.

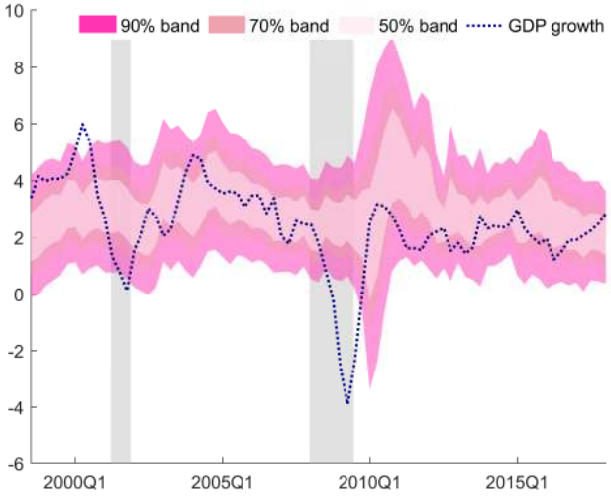
Figure A.5: Predicted probabilities of low growth and low inflation with Azzalini and Capitanio's (2003) skew t distribution



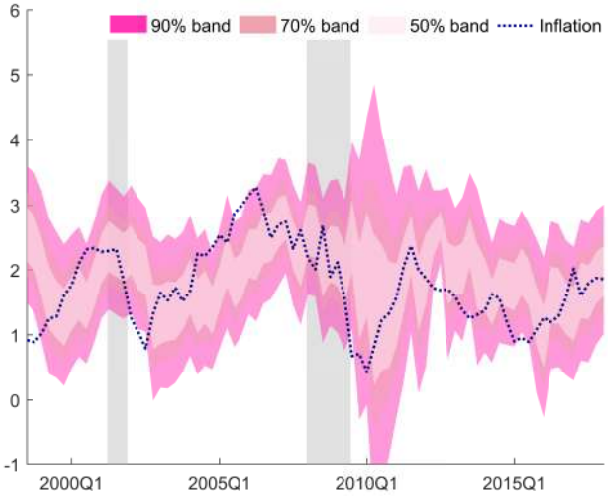
Note: The figure shows according to each model the probabilities of either GDP growth or inflation being less than or equal to 1% (left axis), along with the actual realization of the respective variable (solid blue line, right axis). For an explanation of the different abbreviations, see the main text. The forecast target dates on the horizontal axis range from 1998:Q3 to 2018:Q1. Shaded grey areas denote the periods when the predicted event (e.g. GDP growth ≤ 1%) did in fact occur.

Figure A.6: Predictive intervals of four-quarter-ahead predictive densities of recursively estimated BVAR and CMM

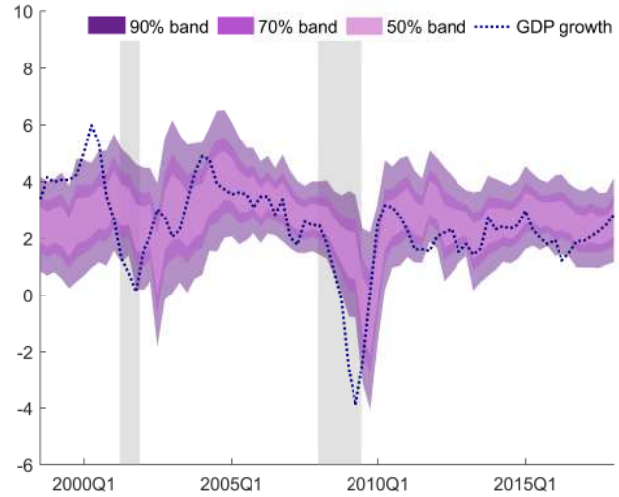
(a) GDP growth, recursive BVAR



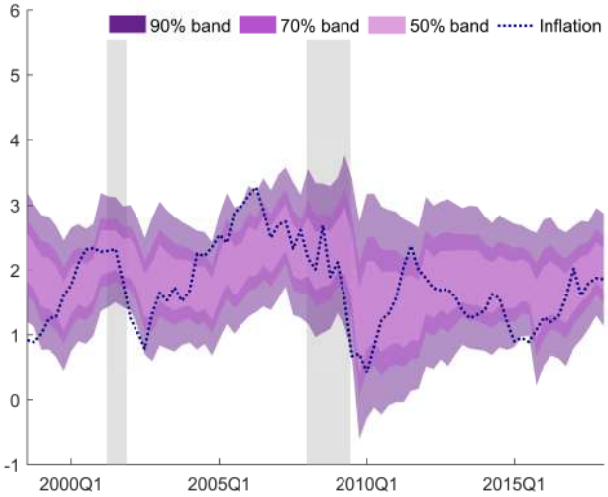
(b) Inflation, recursive BVAR



(c) GDP growth, recursive CMM



(d) Inflation, recursive CMM



Note: The figure shows 90%, 70%, 50% bands, corresponding to the recursively estimated Bayesian VAR's (Panels a and b) and recursively estimated CMM's (Panels c and d) 90%, 70% and 50% equal-tailed predictive intervals for GDP growth and inflation. The dotted blue lines mark the realized values of the variable of interest according to the first release. The forecast target dates on the horizontal axis range from 1998:Q3 to 2018:Q1. Shaded areas are NBER recession periods.

Table A.1: Absolute forecast evaluation: uniformity of PIT

	GDP growth		Inflation	
	KS	CvM	KS	CvM
N	**0.93**(**0.51**)	**0.19**(**0.60**)	**1.16**(**0.27**)	**0.40**(**0.23**)
ST^{JF}	**0.94**(**0.52**)	**0.25**(**0.48**)	**0.91**(**0.48**)	**0.30**(**0.32**)
ST^{AC}	**0.96**(**0.49**)	**0.25**(**0.48**)	**1.02**(**0.36**)	**0.32**(**0.30**)
N (ah)	**0.96**(**0.47**)	**0.26**(**0.46**)	1.68(0.06)	0.90(0.05)
ST^{JF}(ah)	**1.03**(**0.40**)	**0.29**(**0.42**)	1.71(0.05)	0.82(0.06)
ST^{AC}(ah)	**1.03**(**0.40**)	**0.30**(**0.41**)	1.71(0.05)	0.82(0.06)
BVAR (rolling)	2.29(0.00)	1.87(0.00)	**1.27**(**0.28**)	**0.28**(**0.50**)
BVAR (recursive)	**1.47**(**0.12**)	**0.64**(**0.13**)	**1.01**(**0.41**)	**0.24**(**0.45**)
PFE	1.72(0.05)	**0.62**(**0.12**)	**1.45**(**0.18**)	**0.53**(**0.18**)
CMM (rolling)	1.72(0.05)	**0.73**(**0.12**)	**1.44**(**0.17**)	**0.46**(**0.25**)
CMM (recursive)	1.69(0.06)	**0.78**(**0.11**)	**1.40**(**0.19**)	**0.43**(**0.27**)

Note: The table displays the Kolmogorov–Smirnov (KS) and Cramér–von Mises (CvM) test statistics and p-values of the null hypothesis of uniformity of PITs (in parentheses) for different target variables (in the column headers) and models (in rows). For an explanation of the different abbreviations, see the main text. The p-values are calculated using the block weighted bootstrap proposed by Rossi and Sekhposyan (2019), with block length $\ell = 4$ and 10,000 bootstrap replications. The cases in which uniformity cannot be rejected at the 10% level are reported in bold. The survey dates range from 1997:Q4 to 2017:Q2, with corresponding realizations between 1998:Q3 and 2018:Q1.

Table A.2: Absolute forecast evaluation: coverage

	GDP growth		Inflation	
	50%	70%	50%	70%
N	**49.4**(**0.92**)	**67.1**(**0.66**)	**55.7**(**0.36**)	**73.4**(**0.54**)
ST^{JF}	**48.1**(**0.77**)	**63.3**(**0.31**)	**53.2**(**0.61**)	**73.4**(**0.52**)
ST^{AC}	**48.1**(**0.77**)	**63.3**(**0.31**)	**53.2**(**0.61**)	**73.4**(**0.52**)
N (ah)	**41.8**(**0.19**)	57.0(0.06)	63.3(0.04)	81.0(0.03)
ST^{JF}(ah)	**41.8**(**0.19**)	57.0(0.06)	60.8(0.09)	81.0(0.03)
ST^{AC}(ah)	**41.8**(**0.19**)	57.0(0.06)	60.8(0.09)	81.0(0.03)
BVAR (rolling)	**46.8**(**0.62**)	**69.6**(**0.95**)	**40.5**(**0.14**)	55.7(0.03)
BVAR (recursive)	**49.4**(**0.93**)	**67.1**(**0.66**)	**51.9**(**0.78**)	**68.4**(**0.78**)
PFE	65.8(0.02)	**77.2**(**0.22**)	63.3(0.04)	**77.2**(**0.20**)
CMM (rolling)	**49.4**(**0.93**)	**59.5**(**0.12**)	**51.9**(**0.77**)	**69.6**(**0.95**)
CMM (recursive)	**48.1**(**0.78**)	**59.5**(**0.11**)	**51.9**(**0.77**)	**73.4**(**0.56**)

Note: The table displays empirical coverage rates and the two-sided p-values of the null hypothesis that a given coverage rate equals its nominal counterpart (in parentheses) for different target variables at different nominal coverage rates (in the column headers) and models (in rows). For an explanation of the different abbreviations, see the main text. The test statistics were calculated using the Newey and West (1987) HAC estimator with one lag. The cases in which the null hypothesis cannot be rejected at the 10% level are reported in bold. The survey dates range from 1997:Q4 to 2017:Q2, with corresponding realizations between 1998:Q3 and 2018:Q1.

Table A.3: Relative forecast evaluation: CRPS

	GDP growth	Inflation
N	0.75	0.34
ST^{JF}	0.75	0.34
ST^{AC}	0.75	0.54
N (ah)	0.79	0.33
ST^{JF}(ah)	0.79	0.33
ST^{AC}(ah)	0.79	0.34
BVAR (rolling)	0.90	0.45
BVAR (recursive)	0.88	0.39
PFE	0.76	0.32
CMM (rolling)	0.72	**0.32**
CMM (recursive)	**0.72**	0.32
N vs N (ah)	$-0.99(0.16)$	$0.98(0.84)$
ST^{JF} vs ST^{JF}(ah)	$-1.35(0.09)^{*}$	$1.19(0.88)$
ST^{AC} vs ST^{AC}(ah)	$-1.37(0.09)^{*}$	$1.08(0.86)$
N vs BVAR (rolling)	$-1.93(0.03)^{**}$	$-3.22(0.00)^{***}$
N vs BVAR (recursive)	$-1.90(0.03)^{**}$	$-1.76(0.04)^{**}$
ST^{JF} vs BVAR (rolling)	$-2.04(0.02)^{**}$	$-3.14(0.00)^{***}$
ST^{AC} vs BVAR (rolling)	$-1.94(0.03)^{**}$	$0.47(0.68)$
ST^{JF} vs BVAR (recursive)	$-1.96(0.03)^{**}$	$-1.63(0.05)^{*}$
ST^{AC} vs BVAR (recursive)	$-1.82(0.04)^{**}$	$0.79(0.78)$
N vs PFE	$-0.16(0.44)$	$0.74(0.77)$
ST^{JF} vs PFE	$-0.32(0.37)$	$0.94(0.83)$
ST^{AC} vs PFE	$-0.24(0.41)$	$1.12(0.87)$
N vs CMM (rolling)	$0.84(0.80)$	$0.92(0.82)$
N vs CMM (recursive)	$1.07(0.86)$	$0.88(0.81)$
ST^{JF} vs CMM (rolling)	$0.55(0.71)$	$1.12(0.87)$
ST^{AC} vs CMM (rolling)	$0.60(0.73)$	$1.14(0.87)$
ST^{JF} vs CMM (recursive)	$0.76(0.78)$	$1.07(0.86)$
ST^{AC} vs CMM (recursive)	$0.80(0.79)$	$1.13(0.87)$

Note: The target variable used for both estimation and forecast evaluation is shown in the column headers. The top panel displays the Continuous Ranked Probability Score (CRPS) of various density combination methods in the rows. For each variable, the lowest value is in bold. For an explanation of the different abbreviations, please see the main text. The bottom panel displays the Diebold and Mariano (1995) test statistics and p-values (in parentheses, with rejection region in the left tail) comparing predictive accuracy measured by the CRPS. Negative values indicate that the first method outperforms the second one, *, ** and *** denote rejection at the 10%, 5% and 1% significance level, respectively. The test statistics were calculated using the Newey and West (1987) HAC estimator with one lag. The survey dates range from 1997:Q4 to 2017:Q2, with corresponding realizations between 1998:Q3 and 2018:Q1.

Table A.4: Relative forecast evaluation: Brier score

	GDP growth $\leq 1\%$	Inflation $\leq 1\%$
N	0.072	0.124
ST^{JF}	0.073	0.122
ST^{AC}	0.072	0.121
N (ah)	0.076	0.115
ST^{JF}(ah)	0.076	0.113
ST^{AC}(ah)	0.076	0.113
BVAR (rolling)	0.096	0.140
BVAR (recursive)	0.092	0.128
PFE	0.070	0.108
CMM (rolling)	0.063	0.108
CMM (recursive)	**0.062**	**0.107**
N vs N (ah)	$-0.59(0.28)$	$2.63(1.00)$
ST^{JF} vs ST^{JF}(ah)	$-0.55(0.29)$	$2.70(1.00)$
ST^{AC} vs ST^{AC}(ah)	$-0.64(0.26)$	$2.67(1.00)$
N vs BVAR (rolling)	$-1.33(0.09)^{*}$	$-1.62(0.05)^{*}$
N vs BVAR (recursive)	$-1.30(0.09)^{*}$	$-0.75(0.23)$
ST^{JF} vs BVAR (rolling)	$-1.26(0.10)$	$-1.78(0.04)^{**}$
ST^{AC} vs BVAR (rolling)	$-1.30(0.10)^{*}$	$-1.81(0.04)^{**}$
ST^{JF} vs BVAR (recursive)	$-1.22(0.11)$	$-1.04(0.15)$
ST^{AC} vs BVAR (recursive)	$-1.27(0.10)$	$-1.09(0.14)$
N vs PFE	$0.21(0.58)$	$2.37(0.99)$
ST^{JF} vs PFE	$0.36(0.64)$	$2.30(0.99)$
ST^{AC} vs PFE	$0.28(0.61)$	$2.29(0.99)$
N vs CMM (rolling)	$2.13(0.98)$	$3.26(1.00)$
ST^{JF} vs CMM (rolling)	$2.14(0.98)$	$2.92(1.00)$
ST^{AC} vs CMM (rolling)	$2.10(0.98)$	$2.89(1.00)$
N vs CMM (recursive)	$2.28(0.99)$	$3.19(1.00)$
ST^{JF} vs CMM (recursive)	$2.30(0.99)$	$2.93(1.00)$
ST^{AC} vs CMM (recursive)	$2.26(0.99)$	$2.91(1.00)$

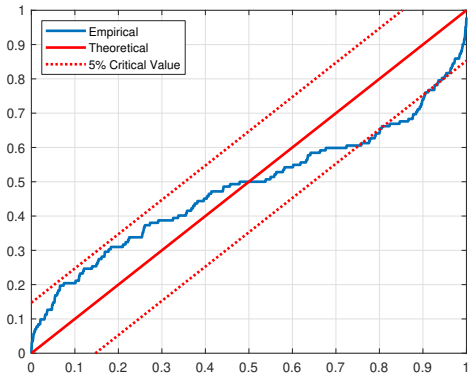
Note: The target variable used for both estimation and forecast evaluation and the corresponding extreme event are shown in the column headers. The top panel displays the Brier score of various density combination methods in the rows. For each variable, the lowest value is in bold. For an explanation of the different abbreviations, please see the main text. The bottom panel displays the Diebold and Mariano (1995) test statistics and p-values (in parentheses, with rejection region in the left tail) comparing predictive accuracy measured by the Brier score. Negative values indicate that the first method outperforms the second one, *, ** and *** denote rejection at the 10%, 5% and 1% significance level, respectively. The test statistics were calculated using the Newey and West (1987) HAC estimator with one lag. The survey dates range from 1997:Q4 to 2017:Q2, with corresponding realizations between 1998:Q3 and 2018:Q1.

Appendix B Further Robustness Results

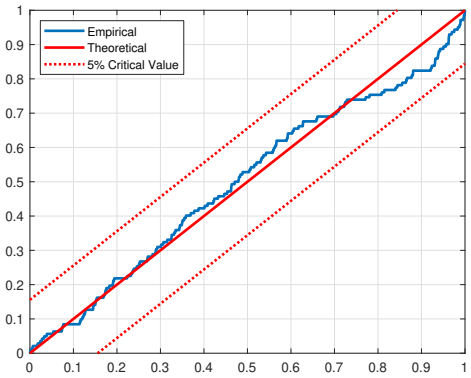
This section shows further robustness studies, where we assess the correct calibration of the component densities in Figure B.1, and an example of local identification of the combination weights in Figure B.2.

Figure B.1: Calibration of component distributions

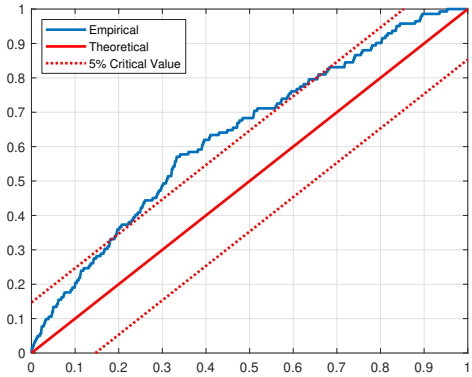
(a) GDP growth, current year



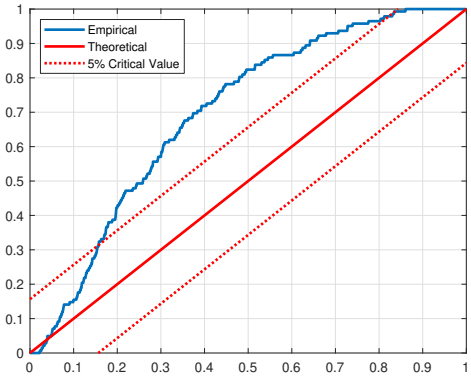
(b) GDP, next year



(c) Inflation, current year

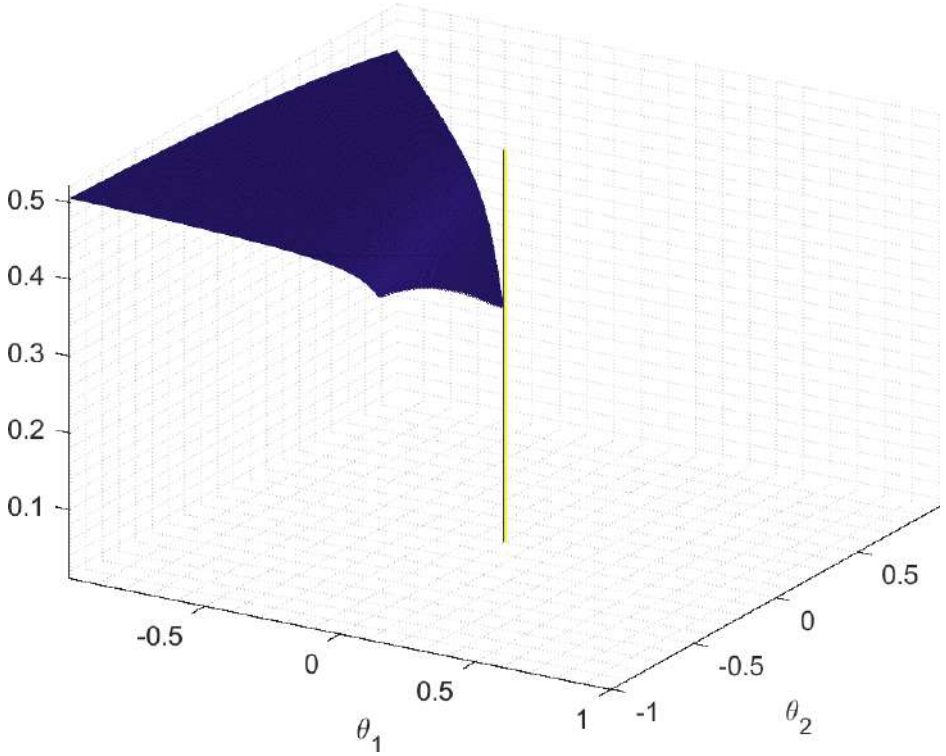


(d) Inflation, next year



Note: The figure shows the empirical CDFs of the current year and next year densities for GDP growth and inflation from the Survey of Professional Forecasters (after fitting the Jones and Faddy, 2003 skew t distribution), the CDF of the PITs under the null hypothesis of correct calibration (the 45 degree line) and the 5% critical values bands based on the Kolmogorov-Smirnov test in Rossi and Sekhposyan (2019), using their bootstrapped critical values.

Figure B.2: Local identification of parameter vector in the 1997:Q4 SPF round, inflation



Note: The figure shows the value of the objective function in Equation (12) (vertical axis) as a function of the parameter vector (θ_1, θ_2) when combining Jones and Faddy (2003) skew t distributions of inflation in the 1997:Q4 SPF round. The vertical spike at $(0.0015, -0.0015)$ marks the optimum.