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]*Central Bank of Hungary and Neumann János University*[*]

[2]*ICREA – Univ. Pompeu Fabra, Barcelona School of Economics, and CREI*[†]

[3]*Texas A&M University*[‡]

September 27, 2021

Abstract

The US Survey of Professional Forecasters produces 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 the density forecasts are mostly 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") or for alternative definitions of the target relative to what is used in the survey. 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 uniformity of the probability integral transform criterion, aiming at obtaining a correctly calibrated fixed-horizon density forecast. We show that our combination method produces competitive density forecasts relative to widely used alternatives based on historical forecast errors or Bayesian VARs. Finally, we elaborate on alternative uses of the proposed methodology to demonstrate the usefulness of our framework for researchers and policymakers.

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

JEL codes: C13, C32, C53.

[*]Senior Researcher (Directorate for International Monetary Policy Analysis) and Assistant Professor (MNB Institute). Address: Szabadsag ter 9, 1054 Budapest, Hungary. Email: ganicsge@mnb.hu.

[†]Professor (Department of Economics). Address: C\Ramon Trias Fargas 25-27, Mercè Rodoreda bldg., 08005 Barcelona, Spain. Email: barbara.rossi@upf.edu.

[‡]Associate Professor (Department of Economics). Address: 4228 TAMU, College Station, TX 77843, USA. Email: tatevik.sekhposyan@gmail.com.

We thank the Editor Kenneth D. West, two anonymous referees, Todd Clark, Marco Del Negro, Jesús Gonzalo, Michael McCracken, Gabriel Pérez Quirós, Minchul Shin, and participants of the 1st Vienna Workshop on Economic Forecasting, VIII$_t$ Zaragoza Workshop in Time Series Econometrics, 2018 Texas Camp Econometrics, 2018 Barcelona GSE Summer Forum on Time Series Econometrics and Applications for Macroeconomics and Finance, 2018 IAAE Annual Conference, 2018 Conference on Real-Time Data Analysis, Methods, and Applications (Philadelphia Fed), 2nd "Forecasting at Central Banks" Conference (Bank of England) and seminars at the Federal Reserve Board, Federal Reserve Bank of San Francisco, Joint Research Centre in Ispra and 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, and Kevin Kliesen for the help with the Summary of Economic Projections data. Part of this research was carried out while Tatevik Sekhposyan was a Visiting Fellow at the Federal Reserve Bank of San Francisco, whose hospitality is greatly acknowledged. The views expressed herein are those of the authors and should not be attributed to the Central Bank of Hungary, Federal Reserve Bank of San Francisco or the Federal Reserve System. Barbara Rossi acknowledges financial support from the Spanish Ministry of Economy and Competitiveness, through the Severo Ochoa Programme for Centres of Excellence in R&D (SEV-2015-0063) and from the Fundación BBVA scientific research grant (PR16_DAT_0043) on Analysis of Big Data in Economics and Empirical Applications.

1 Introduction

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. (2020). Panelists in the US SPF also provide probabilistic forecasts for several variables; for some of them, such as output and prices, even dating back to the first survey round of 1968:Q4. However, these predictions are not widely utilized.

One of the reasons why density forecasts from the US SPF are not used extensively is due to the format of the survey. In each quarter, survey participants communicate their probabilistic forecasts about real GDP growth and inflation (among other variables) for the current and next calendar years via density forecasts. 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. In addition, at times policymakers care about target variables other than what is used in the SPF (for example, year over year growth rates of annual averages in the case of GDP growth). In some cases, such as in the Federal Open Market Committee's quarterly Summary of Economic Projections (SEP), the density forecasts are fixed-event ones, but for fourth quarter over fourth quarter growth rates. Further, it is unclear how to use the SPF densities for this alternative definition of a target variable.

This paper makes two contributions. First, it proposes a density combination approach to obtain fixed-horizon density forecasts from fixed-event ones. We use a novel 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 approach where the weights are assigned to the current and the next year point forecasts proportionally 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 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 forecast error (MSFE) 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 how to best approximate SPF histograms. We investigate the fit of the normal distribution, and extend the empirical literature by considering skew t distributions. The resulting combined (fixed-horizon) density is a mixture, thus flexible, possibly featuring asymmetry, multi-modality and fat tails.

In our paper, we focus on obtaining correctly calibrated fixed-horizon density forecasts. Alternatively, a practitioner can also optimize the weights of the density combination 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. It is important to know that, from a policymaker's point of view, the correctly calibrated 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 note that empirically we might not achieve correct calibration, but we can obtain the "best" calibrated fixed-horizon density forecast — 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. For completeness, however, we investigate how our estimated densities perform relative to commonly used alternatives in the literature.

Our results can be summarized as follows. When using the real GDP growth- and GDP deflator-based inflation density forecasts from the US SPF between 1981:Q3 and 2020:Q1, 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 criteria. 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 (2020) stochastic volatility model using nowcast errors and expectational updates. Furthermore, *on average* there is little gain from fitting the more flexible skew t distributions to the density forecasts provided by the survey histograms as opposed to using normal densities, although the skew t does appear 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.

In terms of decision making, our fixed horizon density forecasts are useful, since in general they are more precise than some of the regularly considered models such as the BVARs or distributions based on historical errors. In addition, our fixed horizon densities can be asymmetric and multi-modal — features that appear to be more pronounced around turning points. These are important for the characterization of the balance of risks at times which are crucial for the economy. A particular episode of such kind is around the financial crisis and the Great Recession in the late 2000s. Our densities emphasize the severity of the situation and need for action, which could be masked when one restricts attention to unimodal densities.

Although our current paper is mainly concerned with constructing four-quarter-ahead year-over-year SPF-based fixed horizon density forecasts for output growth and inflation, our methodology is more general. It can be used to construct densities for quarter-over-quarter growth rates, for other forecast horizons. Overall, our methodology makes SPF density forecast more usable for both policymaking and research. We provide a procedure to obtain timely parameterization of uncertainty while creating a historical dataset that could be used by researchers to use in their models.

The rest of the paper proceeds as follows. Section 2 lays out the proposed methodology. Section 3 discusses the SPF data and the models, while Section 4 presents our results. Section 5 more broadly discusses the usefulness of our framework in other applications. We conclude with Section 6. Additional robustness checks and results are collected in the Online 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.[1]

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 $F^0_{t,q}(\cdot)$ and $F^1_{t,q}(\cdot)$, respectively. 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.[2]

We are interested in constructing a density forecast for a variable that is h quarters ahead of the quarter preceding t, whose CDF is $F^{h,C}_{t,q}(\cdot)$.[3] We estimate this CDF as a convex combination of $F^0_{t,q}(\cdot)$ and $F^1_{t,q}(\cdot)$, 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

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

such that

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

[1] The European Central Bank'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.

[2] 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.

[3] 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.

2.2 Proposed estimator

We propose estimating $\{w_{q,0}^h\}_{q=1}^4$ 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. 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 $y_{t,q}^h$, and forming the PITs. The PIT is the combined CDF evaluated at the realization, formally:

$$\text{PIT}_{t,q}^h \equiv F_{t,q}^{h,C}(y_{t,q}^h) = w_{q,0}^h F_{t,q}^0(y_{t,q}^h) + w_{q,1}^h F_{t,q}^1(y_{t,q}^h)\,. \tag{3}$$

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.

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

$$\Psi_{\mathcal{T}}(r, w_q^h) \equiv |\mathcal{T}|^{-1} \sum_{t \in \mathcal{T}} \mathbb{1}\left[\text{PIT}_{t,q}^h \leq r\right] - r\,, \tag{4}$$

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. However, this would possibly result in considerable estimation uncertainty and make the construction of the out-of-sample density forecasts infeasible.

To accommodate the small sample sizes often encountered in practice, we parametrize the

weights as flexible exponential Almon lag polynomials (Andreou et al., 2010):

$$w^h_{q,0} \equiv \exp(\theta_1 q + \theta_2 q^2)\,, \quad q \in \{1,2,3,4\}\,, \tag{5}$$

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 (5) guarantees that the weights are positive and allows us to pool together PITs from different quarters, using an exponential polynomial.[4] We rely on a rolling window estimation scheme for two reasons. The first is the instabilities in the correct calibration of the SPF density forecasts as documented in Rossi and Sekhposyan (2013), which rolling windows can capture better. In addition, the asymptotic theory for the PIT-based estimator we use is formalized for a rolling window forecasting environment in Ganics (2017). The weights are collected in the vector $w^h \equiv (w^h_{1,0}, w^h_{2,0}, w^h_{3,0}, w^h_{4,0})'$. 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 the scaled quadratic distance:

$$\widehat{w}^h_{q,0} \equiv \exp(\widehat{\theta}_1 q + \widehat{\theta}_2 q^2)\,, \quad q \in \{1,2,3,4\}\,, \tag{6}$$

$$(\widehat{\theta}_1, \widehat{\theta}_2)' \equiv \underset{\theta_1,\theta_2\in\Theta}{\operatorname{argmin}} \int_\rho \frac{\Psi^2_{\mathcal{T}}(r, w^h)}{r(1-r)}\,\mathrm{d}r\,, \tag{7}$$

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.[5] 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.[6] 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

[4] Alternative parametrizations are also known in the literature, e.g. Ghysels et al. (2007) proposed the beta function.

[5] 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 $K = \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \\ 1 & 7 \end{pmatrix}$, $b = (0,0,0,0)'$. 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).

[6] We discovered this pattern when we estimated the weights using the full sample. The restrictions ensure that the real-time weight estimates mimic a similar pattern.

of the PIT and the uniform CDF.[7] In the empirical application, we 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 ensures 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 $F^0_{t,q}(y)$ and $F^1_{t,q}(y)$ distributions to be correctly calibrated, but rather the combination procedure itself serves as a device to achieve the best correctly calibrated combined distribution.

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 approaches 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 2020:Q1. 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,

[7] We use the Anderson–Darling objective function because Ganics (2017) shows via Monte Carlo simulations that it performs better than some of the alternatives, such as Cramér–von Mises- or Kolmogorov–Smirnov-type objective functions. For further details on the estimator, see Ganics (2017).

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.[8] To simplify the notation, we suppress time indices t and q.

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.[9] 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. 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[10] (see

[8] Alternatively, we could use the results in Kheifets and Velasco (2017) to estimate PITs for discrete distributions. We choose our approach because all of the empirical literature we are aware of looking at the SPF density forecasts uses continuous approximations of the histograms.

[9] 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.

[10] An interesting alternative is the generalized beta/triangular distribution used by Engelberg et al. (2009) and Clements (2014). 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

e.g. Giordani and Söderlind, 2003, D'Amico and Orphanides, 2008, Clements, 2014 or Rossi et al., 2017), but to the best of our knowledge, we are the first to use the aforementioned skew t distribution to model survey forecasts. The use of the skew t distribution is motivated by the observation that, as we will illustrate later on, the survey forecast histograms seem to be better approximated by a skewed underlying distribution in several cases. The JF distribution generalizes Student's t distribution by introducing the parameters $a, b > 0$ regulating skewness and tail behavior at the same time. It 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 PDF at $x \in \mathbb{R}$ is given by

$$f(x;\mu,\sigma,a,b) = \frac{1}{\sigma} C_{a,b}^{-1} \left(1+\tau\right)^{a+1/2} \left(1-\tau\right)^{b+1/2}, \tag{8}$$

$$C_{a,b} = 2^{a+b-1} B(a,b)(a+b)^{\frac{1}{2}}, \tag{9}$$

$$\tau = \frac{x-\mu}{\sigma}\left(a+b+\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{1}{2}}, \tag{10}$$

while the CDF is given by

$$F(x;\mu,\sigma,a,b) = I_z(a,b), \tag{11}$$

$$z = \frac{1}{2}\left(1+\frac{\left(\frac{x-\mu}{\sigma}\right)}{\sqrt{a+b+\left(\frac{x-\mu}{\sigma}\right)^2}}\right), \tag{12}$$

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).[11]

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 $\{s_i\}_{i=1}^S$

no need to close the tail bins of the histograms in an arbitrary manner.

[11] Azzalini and Capitanio (2003) proposed an alternative skew t distribution. When using that distribution, our results are largely unchanged, demonstrated in Online Appendix B. The JF distribution is appealing as its CDF can be evaluated quickly, while 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.

associated with the intervals/bins specified by the survey, in the spirit of Engelberg et al. (2009):

$$\widehat{\theta}_d = \underset{\theta_d \in \Theta_d}{\operatorname{argmin}} \sum_{i=1}^{S} \left(F_d(s_i; \theta_d) - F(s_i) \right)^2 . \tag{13}$$

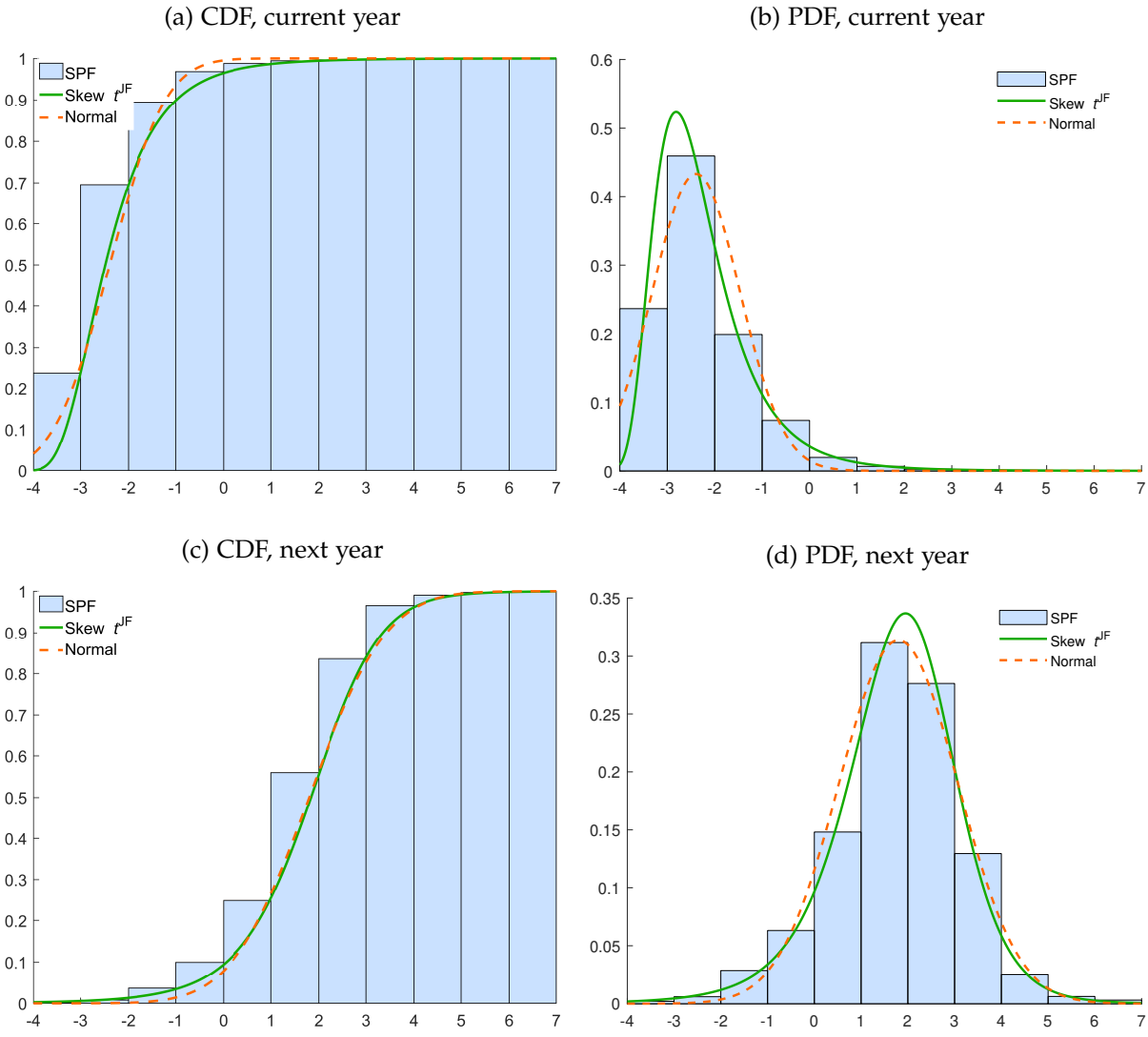
By estimating the distributions in each quarter, the procedure described above gives us sequences of predictive CDFs $F^0_{t,q,d}(y)$ and $F^1_{t,q,d}(y)$ for each variable of interest. Analogously, the corresponding predictive PDFs are denoted by $f^0_{t,q,d}(y)$ and $f^1_{t,q,d}(y)$. For example, take 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 Skew t^{JF}). The fitted distributions range from the 1981:Q3 survey round to the 2020:Q1 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 2020:Q1 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 Philadelphia Fed'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, which we take as the point of reference. Hence, 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 2020:Q4. While it is not unequivocally clear that SPF respondents target the first release, they tend to more accurately forecast the earlier releases (see Stark, 2010), which motivated our choice of the target variable.

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

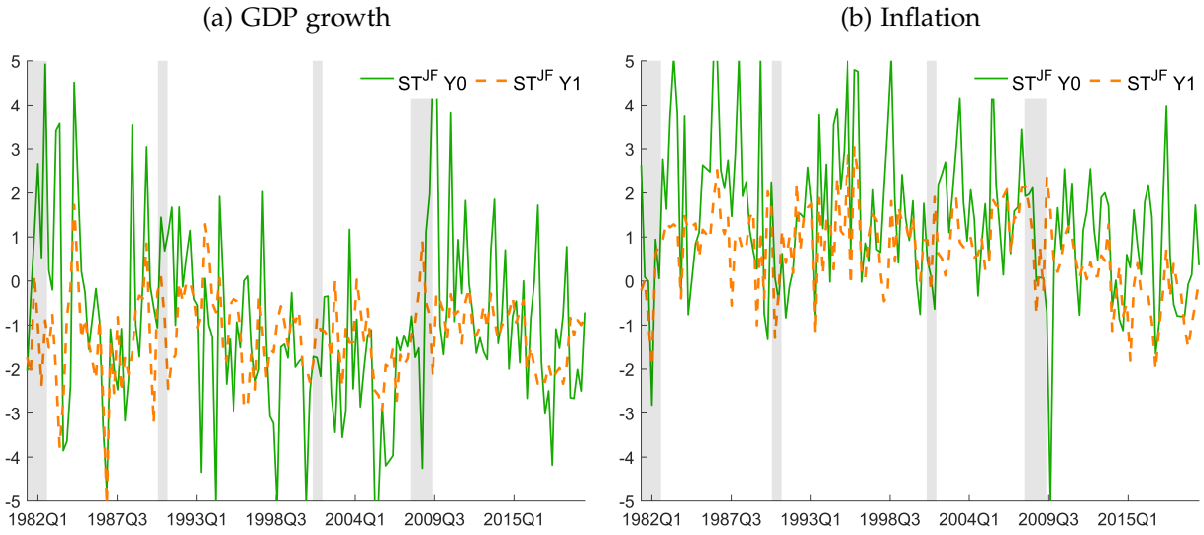


Note: The figures show the empirical CDFs and histograms of the SPF forecasts of GDP growth for 2009 and 2010, according to the 2009:Q2 survey round, and the fitted skew t (solid green curve) and normal (dashed orange curve) distributions' CDFs and PDFs.

We estimate the combination weights in rolling windows of $R = 60$ quarters using the Anderson–Darling-type objective function in Equation (7), with $\rho = [0, 1]$. For example, take the case of GDP growth. The first out-of-sample prediction, corresponding to 1998:Q3, is obtained as follows. To estimate the combination weights in Equation (6), we use the fitted distributions (either normal or JF) between the 1981:Q3 and 1996:Q4 survey rounds, and the GDP growth realizations between 1982:Q2 and 1997:Q3. These weights would have been available in 1997:Q4 at the earliest due to the publication lag of the 1997:Q3 GDP figure. We combine the fixed-event predictive distributions of the 1997:Q4 survey round using these combination weights $(\widehat{w}^4_{4,0}, \widehat{w}^4_{4,1})'$.[12] We repeat this estimation procedure in a rolling window

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

Figure 2: Skewness of fitted skew t distributions between 1981:Q3 and 2020:Q1



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.

manner until the end of the sample, where the weights are used to combine densities from the 2020:Q1 survey round, predicting GDP growth for 2020:Q4. As a result, we obtain fixed-horizon density estimates in real-time.

In the next section, the models using the normal or the skew t distribution are denoted by N or ST^{JF}, respectively. We compare our suggested weight estimation procedure against the deterministic, fixed weights given in Equation (14). When the deterministic mixture weights are used, we add "(det)" after the distribution's abbreviation, while the models using estimated weights are not accompanied by additional notation. It should be noted that both the estimated as well as the deterministic 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 latter weights come with the advantage that they do not need to be estimated, the disadvantage being that it is not obvious how to derive such weights for general cases (see Footnote 13).

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.

3.3.1 Deterministic weight combinations

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 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 one-year-ahead ($h = 4$) quarterly *density* forecasts. Based on this approach, in quarter one, the current year density receives 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

$$w^4_{q,0} = \frac{5-q}{4}, \qquad w^4_{q,1} = \frac{q-1}{4}, \quad q \in \{1,2,3,4\}. \tag{14}$$

One advantage of Dovern et al.'s (2012) method is that the weights are given and need not be estimated, which is useful when the sample is too short to allow estimation. However, it is unclear whether an analogous combination scheme is applicable in situations with different forecast horizons or frequencies.[13] Furthermore, these weights 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 and the MSFE criterion only.

3.3.2 BVAR with stochastic volatility

The second benchmark model is a BVAR with stochastic volatility based on Clark and Ravazzolo (2015), which 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, and is re-estimated in rolling windows of 60 quarterly observations. The raw GDP and GDP deflator data are obtained as before, the 3-month Treasury bill rate is 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 from the FRED database (mnemonic: UNRATE).

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

In the real-time spirit of the forecasts, the estimation procedure only uses data available up to the previous quarter at each forecast origin. The first rolling window BVAR forecasts of four-quarter-ahead GDP growth and inflation are for 1998:Q3. We use the same prior specification as Clark and Ravazzolo (2015), except that the parameters of the prior distribution are re-estimated in each rolling window. We simulate 80,000 draws (after 20,000 burn-in) from the 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 these predictive distributions to obtain distributions for year-on-year growth rates.

3.3.3 The Past Forecast Error (PFE) model

As a third benchmark model, we estimate predictive distributions based on past forecast errors (PFE) using the SPF point forecasts. Clements (2018) has shown that this approach performs well, 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.

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.[14] 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. (2020) use a similar benchmark and show that it performs well, particularly in the sample period considered in our analysis.

3.3.4 Clark et al.'s (2020) (CMM) model

Our final benchmark is the model recently proposed by Clark et al. (2020), who use 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

[14]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.

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 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. 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 p.23 of Clark et al., 2020), 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 Markov Chain Monte Carlo output, resulting in 10,000 draws used in our analysis. For further details on the model and the estimation procedure, see Clark et al. (2020).

3.3.5 Relationship between our approach and the benchmark approaches

It is important to note that the benchmark 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 (quarter-on-quarter) CMM and the PFE predictive distributions are unimodal and symmetric by construction, while our more general mixture densities can display multi-modality and asymmetry due to the properties of the component distributions as opposed to arising from parameter estimation error only. 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.

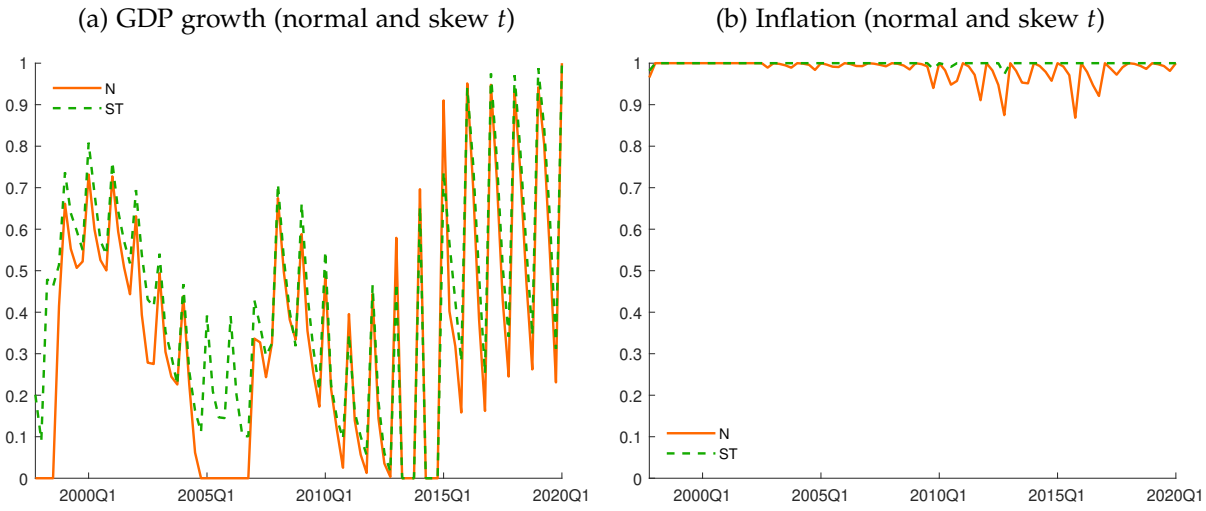
4 Empirical Results

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

Figure 3 shows the weights on current year's distribution that are used for density combination in each quarter. Recall that in each SPF round, Equations (6) and (7) provide weight

estimates $\widehat{w}^4_{q,0}$ for $q = 1, \ldots, 4$, displayed in Figure A.1 in Online Appendix A. 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 Panel a that the estimated weights display substantial time-variation, and seasonality, with local peaks (troughs) in the first (fourth) quarter of each year. Panel b, instead, displays an entirely different pattern for inflation, where 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. The seasonal pattern seen in the case of GDP growth is due to the fact that the information content of current year's forecast decreases as one goes from one quarter to the next. Intuitively, as one progresses through a calendar year, next year's density forecast should receive a larger weight as long as it is not so miscalibrated that including it with a positive weight would deteriorate the calibration of the combined density. This happens for inflation, where next year's density is heavily miscalibrated, shown in Online Appendix C.[15]

Figure 3: Weights on current year's density forecast



Note: The panels in the figure depict the estimated combination weights one would apply in each SPF round (horizontal axis) to current year's density to combine the forecasts.

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

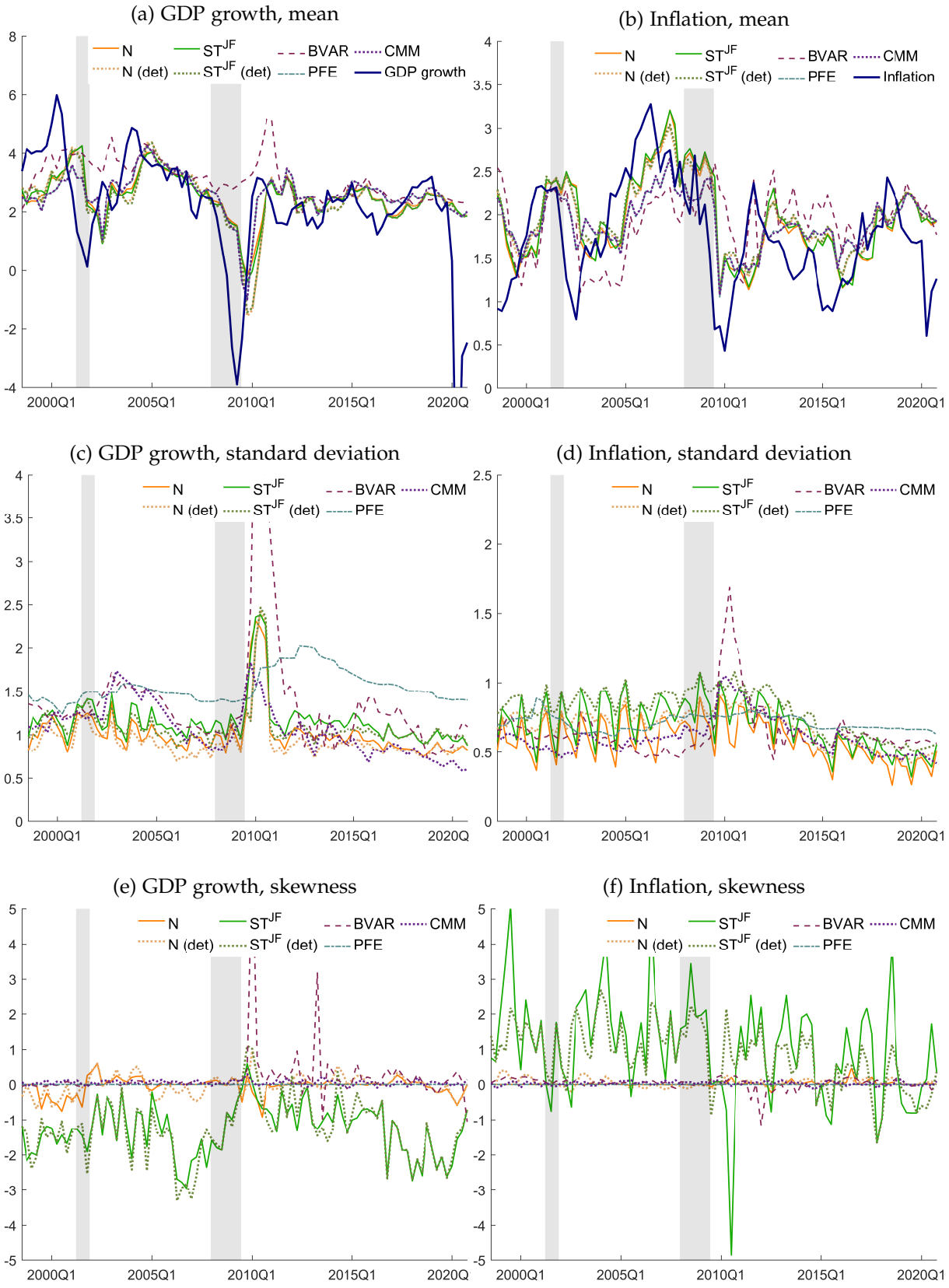
[15]We have investigated whether this behavior of weights for inflation is due to the lack of identification of the weights. As it turns out, in general, the objective function is not flat in the neighborhood of the estimated parameters, suggesting at least local identification (see an example in Figure C.2).

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 Great 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 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. On the other hand, the CMM model and the BVAR display very little skewness, apart from two brief instances in the case of the BVAR, shown in Panel e.

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 realization, regardless of how the underlying densities are approximated. Moreover, based on the skew t mixtures, there is a higher probability associated with the "bad" mode. 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. Adrian et al. (2021) also document bimodality in the conditional (on financial conditions) predictive distribution of the GDP growth around the Great Recession.

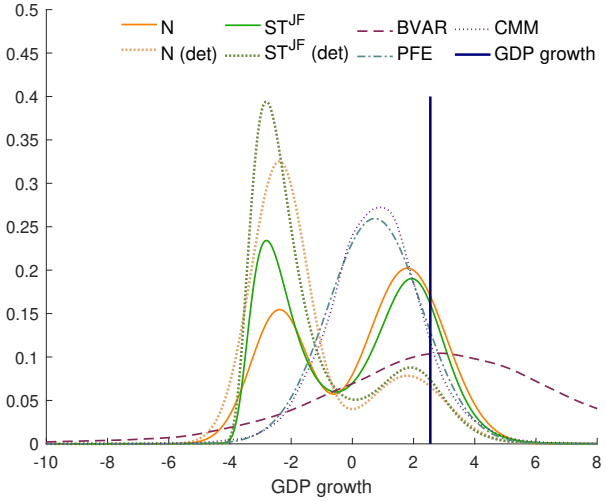
The bimodality discussed above stems from our procedure — we are mixing densities and bimodality will occur when these densities are considerably different from each other. This certainly does not imply that the consensus survey forecasts would have been multi-modal if predictions about fixed-event (four-quarter-ahead) densities had been elicited directly. In fact, our objective is to obtain a correctly calibrated fixed-horizon density: multi-modality may be optimal in this case, particularly, when we have a constrained information set in terms of input densities. However, multi-modality, at least in the sample we consider, is a rare and short-lived event and does not heavily influence the *average* performance of the models.

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



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 2020:Q4. 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

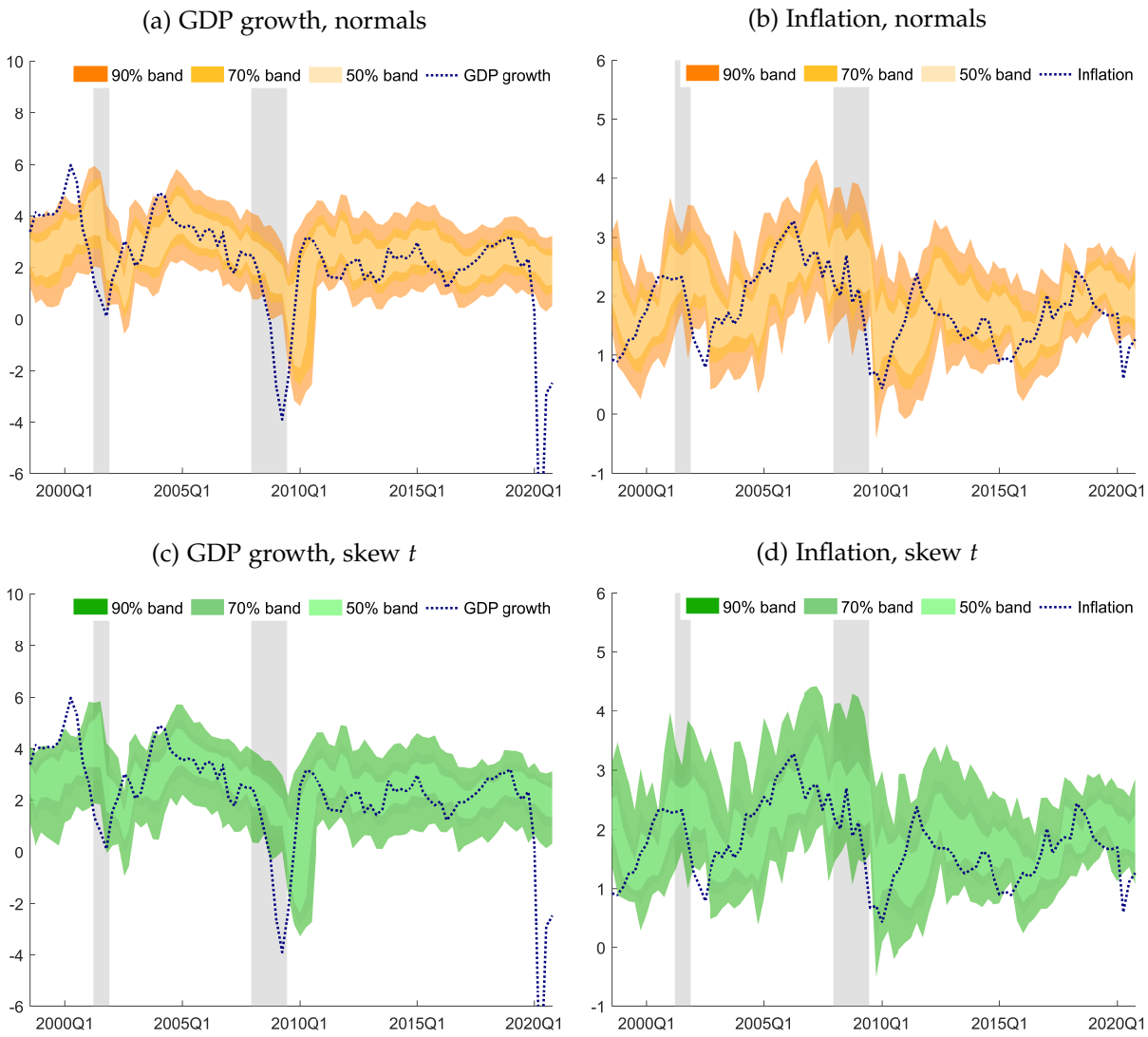


Note: The figure shows the four-quarter-ahead predictive densities of various models for GDP growth, as of 2009:Q2. The solid vertical line indicates the actual realization of GDP growth in 2010:Q1. For an explanation of the different abbreviations, see the main text.

Figure 6 shows various quantiles associated with our combined density for inflation and output growth. For instance, the 90% interval in the figures is defined as the interval between the 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. For analogous figures using the benchmark models, see Online Appendix A.

The panels in Figure 6 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 (see Online Appendix A) 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 HDRs are the intervals of shortest length with a given target coverage, say 90%. When the distributions are unimodal and symmetric, these two measures are the same. To see whether asymmetry and multi-modality are important in some periods, Figure 7 shows the 50%, 70% and 90% HDRs, calculated using the density quantile approach of Hyndman (1996).

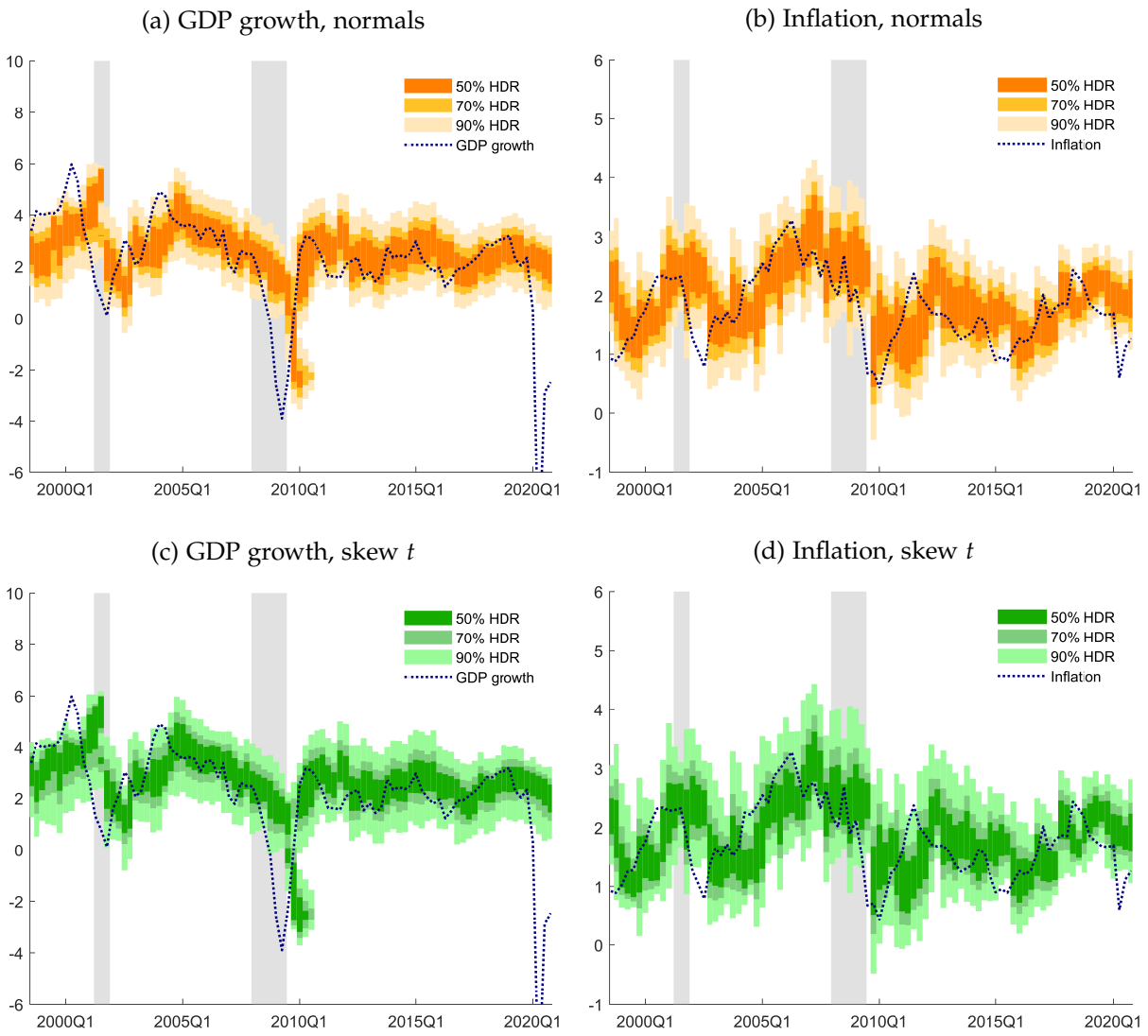
Figure 6: Predictive intervals of four-quarter-ahead combined predictive densities, using estimated weights



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 2020:Q4. Shaded areas denote NBER recession periods.

When we compare Figure 6 to Figure 7, 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 does not display any disjoint intervals, communicating unimodality. This result is not surprising since our density forecast approach almost exclusively puts all the weight on the current year density (the best calibrated density

Figure 7: 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 2020:Q4. Shaded areas denote NBER recession periods.

among the inputs). Consequently, the resulting density is not a mixture, but it is mainly the consensus density associated with current year inflation. 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 minimized by the mode of the distribution). In the case of the BVAR and CMM models, we find no major differences between the equal-tailed intervals and HDRs (see Online Appendix A).

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.[16] Table 1 shows the results. Each cell displays the p-value of testing the null hypothesis of the uniformity of the PIT against the alternative of miscalibration. 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 deterministic 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.

[Table 1 about here.]

To gain a better understanding of the absolute performance of the various density forecast approaches, we report the empirical coverage rates at the 50%, 70% and 90% nominal rates. For each variable and forecasting method, we determined e.g. the 25th and 75th percentiles (50% nominal rate) of the predictive distributions in each quarter (as displayed in Figure 6, and Figures A.3 and A.4 in Online Appendix A), and calculated the ratio of cases when the realization of a particular variable fell inside these intervals. Then, we performed a two-sided t-test to test the null hypothesis that a given coverage rate equals its nominal counterpart.[17]

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 deterministic mixtures for GDP growth display significant undercoverage. For GDP growth, both the BVAR and the CMM model deliver correct coverage (except at the 70% level in the case of the latter), while the PFE model significantly overcovers at the 50% rate. On the other hand, for inflation, only the CMM model displays correct coverage at all levels but not the BVAR or the PFE models.

[Table 2 about here.]

[16] We use the block weighted bootstrap proposed by Rossi and Sekhposyan (2019), with block length $\ell = 4$ and 10,000 bootstrap replications to take into account the serial correlation associated with multi-step-ahead PITs.

[17] The asymptotic variance is calculated using the Newey and West (1987) heteroskedasticity and autocorrelation consistent (HAC) estimator with one lag. The sequences of 0–1 indicators corresponding to realizations outside and inside the predictive intervals display low serial correlation.

The Continuous Ranked Probability Score (CRPS) has been used in several studies to evaluate competing forecasts (e.g. Clark et al., 2020). Formally, for the h-quarter-ahead density forecast made in year t and quarter q using model m, it is defined as

$$\text{CRPS}_{t,q}^{h,(m)} \equiv \int_{-\infty}^{\infty} \left(F_{t,q}^{h,(m)}(y) - \mathbb{1}\left[y_{t,q}^{h} \leq y \right] \right)^2 \mathrm{d}y \,, \tag{15}$$

where $F_{t,q}^{h,(m)}(y)$ is the predictive CDF. The average full-sample CRPS is given by

$$\text{CRPS}^{h,(m)} \equiv |\mathcal{T}|^{-1} \sum_{t \in \mathcal{T}} \text{CRPS}_{t,q}^{h,(m)} \,. \tag{16}$$

Lower values of the CRPS correspond to better models. For the mixture densities, we numerically calculate the integral in Equation (15), while for the MCMC-based densities, such as those obtained from a BVAR and CMM, we used the empirical CDF-based approximation proposed by Krüger et al. (2017). Following Gneiting and Raftery (2007), we can analytically evaluate the CRPS for the PFE since the predictive density is normal.

The top panel in Table 3 shows the CRPS of the proposed density combination, along with various benchmarks. The bottom panel in Table 3 reports the Diebold and Mariano (1995) and West (1996) test statistics and the corresponding p-values (in parentheses) when equal predictive ability is measured by the CRPS against a one-sided alternative. 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.

[Table 3 about here.]

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 deterministic counterparts, and they significantly outperform the BVAR at the 5% significance level. When predicting inflation, our method beats the BVAR, while the deterministic 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 researchers and policymakers. To evaluate how each model performs in forecasting extreme events, we performed the following exercise. Using each model, we calculated 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 inflation. Figure 8 shows the probabilities for each event (scaling on the left axis), along with the actual realizations of the variable (scaling on the right axis).[18] 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 its 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 8 (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 mixtures.

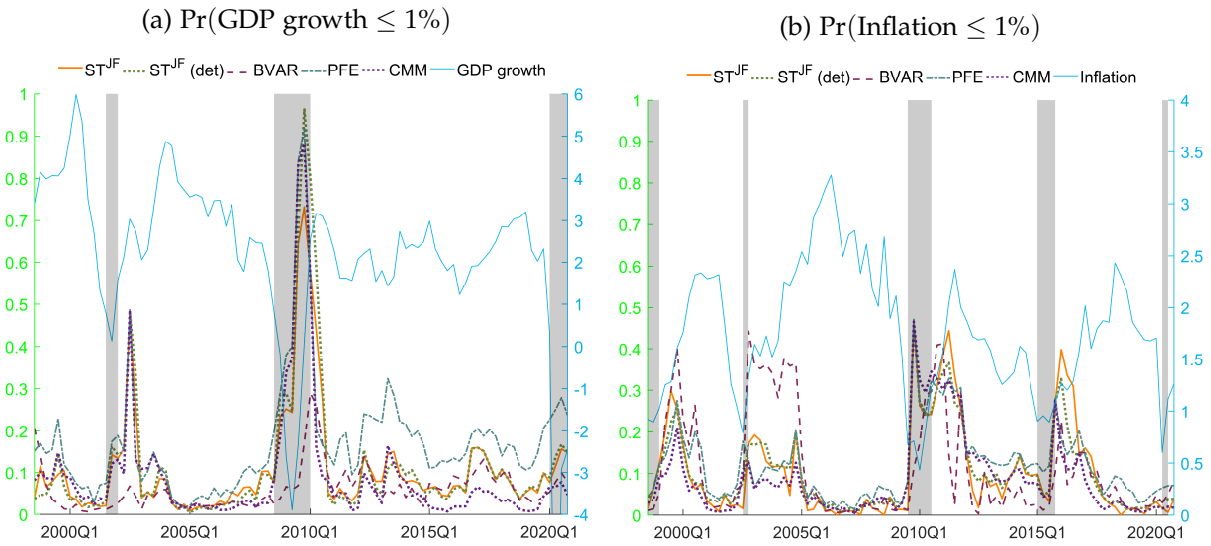
We formally evaluated each model's predictions for the aforementioned extreme events using the Brier (or quadratic) score (Gneiting and Raftery, 2007). For the h-quarter-ahead density forecast made in year t and quarter q using model m and at threshold k (in our case, $k = 1\%$), it is defined as

$$\mathrm{BS}_{t,q}^{h,(m)}(k) \equiv \left(F_{t,q}^{h,(m)}(k) - \mathbb{1}\left[y_{t,q}^{h} \leq k\right]\right)^2, \tag{17}$$

which is precisely the integrand in Equation (15). Lower values of the Brier score correspond

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

Figure 8: 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 2020:Q4. Shaded grey areas denote the periods when the predicted event (e.g. GDP growth $\leq 1\%$) did in fact occur.

to better predictions. The full-sample Brier score is defined analogously as

$$\mathrm{BS}^{h,(m)}(k) \equiv |\mathcal{T}|^{-1} \sum_{t \in \mathcal{T}} \mathrm{BS}_{t,q}^{h,(m)}(k)\,. \tag{18}$$

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) and West (1996) 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 deterministic 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 8 shows that these were indeed very rare events. We expect that the predictability of tail events could be improved if weights are estimated targeting correct calibration in the tails.

[Table 4 about here.]

5 Alternative Uses of the Proposed Method

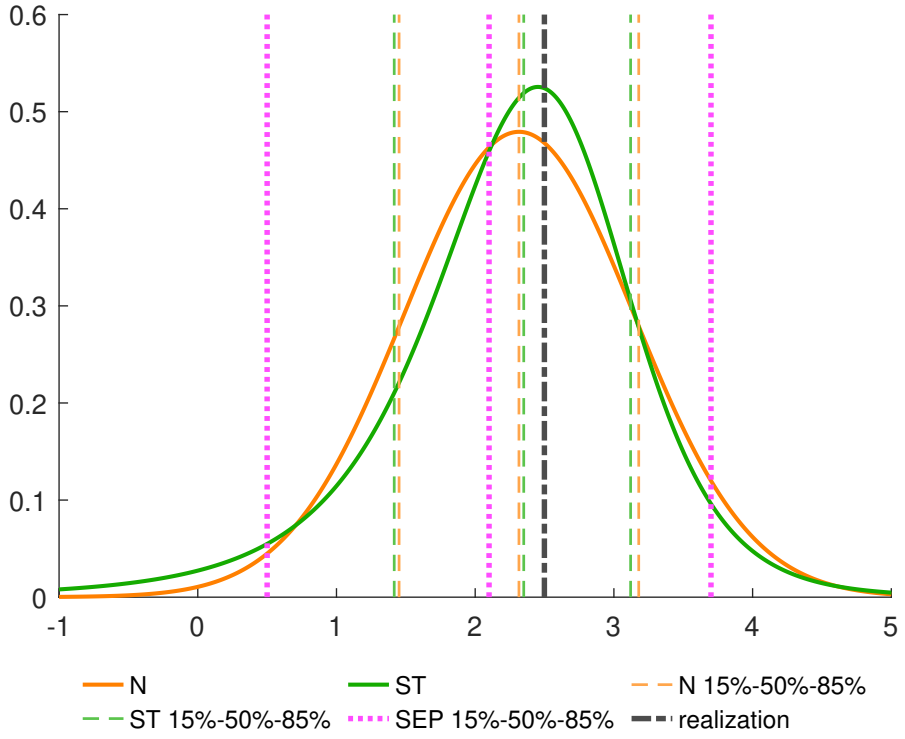
We have proposed a methodology that uses the information in the fixed-horizon SPF density forecasts to obtain four-quarter-ahead density forecasts. In this section, we highlight briefly that while in our view this is an important use of the methodology, as it provides a timely description of future uncertainty, but not the only one. As mentioned in the Introduction, there are a number of ways our method can be used.

First, one could use our methodology to obtain fixed-event density forecasts, but for other targets. For instance, in the SEP, the Federal Reserve produces fixed-event forecasts, but for fourth-quarter over fourth-quarter growth rates as opposed to annual average over annual average growth rates.[19] Figure 9 contrasts our combined-density-implied intervals with those reported in the SEP. Given that the SPF is a quarterly survey, we use the survey round closest to the Federal Open Market Committee meeting date for which the SEPs were prepared. The SPF densities we use have a survey deadline date of February 7, 2017. The SEP intervals are for the FOMC meeting of March 14-15, 2017.

In this example, our densities are skewed to the left, regardless of whether the underlying histograms are proxied with a normal or a skew t distribution. Moreover, they are centered on a higher mean value, which is closer to the realization. In addition, there is less uncertainty about GDP growth since the 70% predictive intervals are tighter. Overall, it appears that the density we obtain is shifted to the right relative to the SEP density. The SEP also includes an auxiliary question about GDP uncertainty to gauge whether the uncertainty perceived by each FOMC member is lower, broadly similar, or higher than that reported in the figure. During this particular meeting, 12 participants reported broadly similar perceived uncertainty, while four participants reported higher and no participant reported lower uncertainty. Overall, our combined densities suggest lower level of uncertainty than reported in the SEP and perceived by the FOMC participants.

[19] Data on SEP densities are available at https://www.federalreserve.gov/monetarypolicy/fomcminutes20170315ep.htm.

Figure 9: Density combination and SEP densities for GDP growth in 2017:Q1



Note: The figure shows our combined densities where we combine SPF histograms, approximated either by a normal (N) or a skew t distribution (ST). Importantly, the target variable is a Q4 over Q4 growth rate of GDP (realization). We provide the 70% equal-tailed predictive intervals and contrast it to what is reported in the SEP.

This particular application is important since, for instance, in June of 2020, the SEP did not contain predictive intervals given that the historical forecast errors would not be representative of the uncertainty in times of a pandemic. In fact, the Federal Reserve did not communicate predictive intervals in the first half of 2020. On the other hand, the SPF still provides density forecasts, and we can still produce predictive densities to complement the policy analysis.

Second, a researcher can choose to minimize the distance between the CDF of the combined distribution and some other distribution. 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. A researcher can address this issue by considering 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). This approach will provide quarterly time series of one-step-ahead density forecasts, which could be more useful than the annual frequency data

that is typically used in the literature (see Andrade et al., 2016 for a sub-sampling exercise in the context of point forecasts provided by the Blue Chip Financial Forecasts).

Finally, while one might think that an easy solution to the lack of fixed-horizon density forecasts or predictive distributions for alternative target variables is to modify the survey, the issue would be that variables can only be added going forward and not backward (for example, in 2007 the SPF was extended by fourth-quarter over fourth-quarter CPI and PCE inflation probability forecasts). These surveys are used not only for characterizing uncertainty into the future, but also for disciplining economic models, in which case, a long time series is crucial. For example, it is often of interest to use historical time series of survey density forecasts to discipline learning dynamics, as in Chatterjee and Milani, 2019. Our framework provides a viable solution for this situation.

6 Conclusion

This paper proposes a density combination methodology to extract information from fixed-event densities to construct correctly calibrated fixed-horizon density forecasts. 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 our procedure achieves correct calibration for predictive densities in a real-time out-of-sample exercise. 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 Bayesian VAR with stochastic volatility. Since our predictive distributions are based on survey information, they are timely. Moreover, they are precise and due to the flexible mixture specification, they are informative about the balance of risks particularly at turning points. Our approach makes fixed-event density forecasts more useful for research, policy analysis and communication.

References

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

Adrian, T., Boyarchenko, N., and Giannone, D. (2021). Multimodality in Macro-Financial Dynamics. *International Economic Review*, 62(2):861–886.

Andrade, P., Crump, R. K., Eusepi, S., and Moench, E. (2016). Fundamental disagreement. *Journal of Monetary Economics*, 83:106–128.

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.

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

Chatterjee, P. and Milani, F. (2019). Perceived Uncertainty Shocks, Excess Optimism-Pessimism, and Learning in the Business Cycle. Technical report.

Clark, T. E., McCracken, M., and Mertens, E. (2020). Modeling Time-Varying Uncertainty of Multiple-Horizon Forecast Errors. *Review of Economics and Statistics*, 102(1):17–33.

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. (2014). 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.

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):253–263.

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.

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.

Gneiting, T. and Ranjan, R. (2011). Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics*, 29(3):411–422.

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.

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.

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. (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.

Stark, T. (2010). Realistic Evaluation of Real-Time Forecasts in the Survey of Professional Forecasters. Research Rap, Special Report, Federal Reserve Bank of Philadelphia.

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.

West, K. D. (1996). Asymptotic Inference about Predictive Ability. *Econometrica*, 64(5):1067–1084.

Table 1: Absolute forecast evaluation: uniformity of PIT

	GDP growth		Inflation	
	KS	CvM	KS	CvM
N	**0.54**	**0.46**	**0.17**	**0.12**
ST	**0.41**	**0.36**	**0.18**	**0.16**
N (det)	**0.32**	**0.34**	0.03	0.02
ST (det)	**0.31**	**0.32**	0.03	0.03
BVAR	0.00	0.00	**0.18**	**0.30**
PFE	0.05	0.08	**0.12**	0.09
CMM	0.09	0.09	**0.16**	**0.13**

Note: The table displays the p-values of Kolmogorov–Smirnov (KS) and Cramér–von Mises (CvM) tests of the null hypothesis of uniformity of PITs for different target variables (in the column headers) and models (in rows). N and ST correspond to the combinations of normal and skew t distributions using our proposed weight estimates (see Section 3.2), while N (det) and ST (det) denote their conterparts using deterministic weights (see Section 3.3.1). PFE corresponds to the normal distribution based on past forecast errors (Section 3.3.3), BVAR is the Bayesian VAR of Section 3.3.2, while CMM is the stochastic volatility model based on point forecast revisions (Section 3.3.4). 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 2020:Q1, with corresponding realizations between 1998:Q3 and 2020:Q4.

Table 2: Absolute forecast evaluation: coverage

	GDP growth			Inflation		
	50%	70%	90%	50%	70%	90%
N	**46.7**(**0.58**)	**66.7**(**0.59**)	76.7(0.01)	**51.1**(**0.85**)	**68.9**(**0.84**)	**85.6**(**0.27**)
ST	**45.6**(**0.46**)	**63.3**(**0.29**)	81.1(0.07)	**48.9**(**0.85**)	**68.9**(**0.84**)	**86.7**(**0.42**)
N (det)	40.0(0.08)	57.8(0.06)	76.7(0.01)	**57.8**(**0.20**)	**76.7**(**0.20**)	**91.1**(**0.75**)
ST (det)	40.0(0.08)	57.8(0.06)	78.9(0.03)	**55.6**(**0.36**)	**76.7**(**0.20**)	**91.1**(**0.73**)
BVAR	**47.7**(**0.59**)	**68.9**(**0.84**)	**85.6**(**0.35**)	40.0(0.09)	57.8(0.05)	77.8(0.02)
PFE	65.6(0.01)	**75.6**(**0.33**)	**91.1**(**0.77**)	61.1(0.07)	**76.7**(**0.22**)	95.6(0.02)
CMM	**47.8**(**0.73**)	58.9(0.08)	**82.2**(**0.10**)	**48.9**(**0.86**)	**70.0**(**1.00**)	**87.8**(**0.57**)

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 the definition of the model abbreviations, see Table 1. 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 2020:Q1, with corresponding realizations between 1998:Q3 and 2020:Q4.

Table 3: Relative forecast evaluation: CRPS

	GDP growth	Inflation
N	0.92	0.35
ST^{JF}	0.91	0.35
N (det)	0.95	0.34
ST^{JF}(det)	0.94	0.34
BVAR	1.06	0.44
PFE	0.91	0.33
CMM	**0.89**	**0.33**
N vs N (det)	$-0.96(0.17)$	$1.24(0.89)$
ST vs ST (det)	$-1.24(0.11)$	$1.46(0.93)$
N vs BVAR	$-2.07(0.02)^{**}$	$-3.01(0.00)^{***}$
ST vs BVAR	$-2.19(0.01)^{**}$	$-2.92(0.00)^{***}$
N vs PFE	$0.16(0.56)$	$0.92(0.82)$
ST vs PFE	$-0.06(0.48)$	$1.14(0.87)$
N vs CMM	$0.91(0.82)$	$1.00(0.84)$
ST vs CMM	$0.58(0.72)$	$1.21(0.89)$

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 the definition of the model abbreviations, see Table 1. The bottom panel displays the Diebold and Mariano (1995) and West (1996) 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 2020:Q1, with corresponding realizations between 1998:Q3 and 2020:Q4.

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 (det)	0.076	0.115
ST^{JF}(det)	0.076	0.113
BVAR	0.096	0.140
PFE	0.070	0.108
CMM	**0.063**	**0.108**
N vs N (det)	$-0.59(0.28)$	$2.63(1.00)$
ST^{JF} vs ST^{JF}(det)	$-0.55(0.29)$	$2.70(1.00)$
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 the definition of the model abbreviations, see Table 1. The bottom panel displays the Diebold and Mariano (1995) and West (1996) 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 2020:Q1, with corresponding realizations between 1998:Q3 and 2020:Q4.

Online Appendix to

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

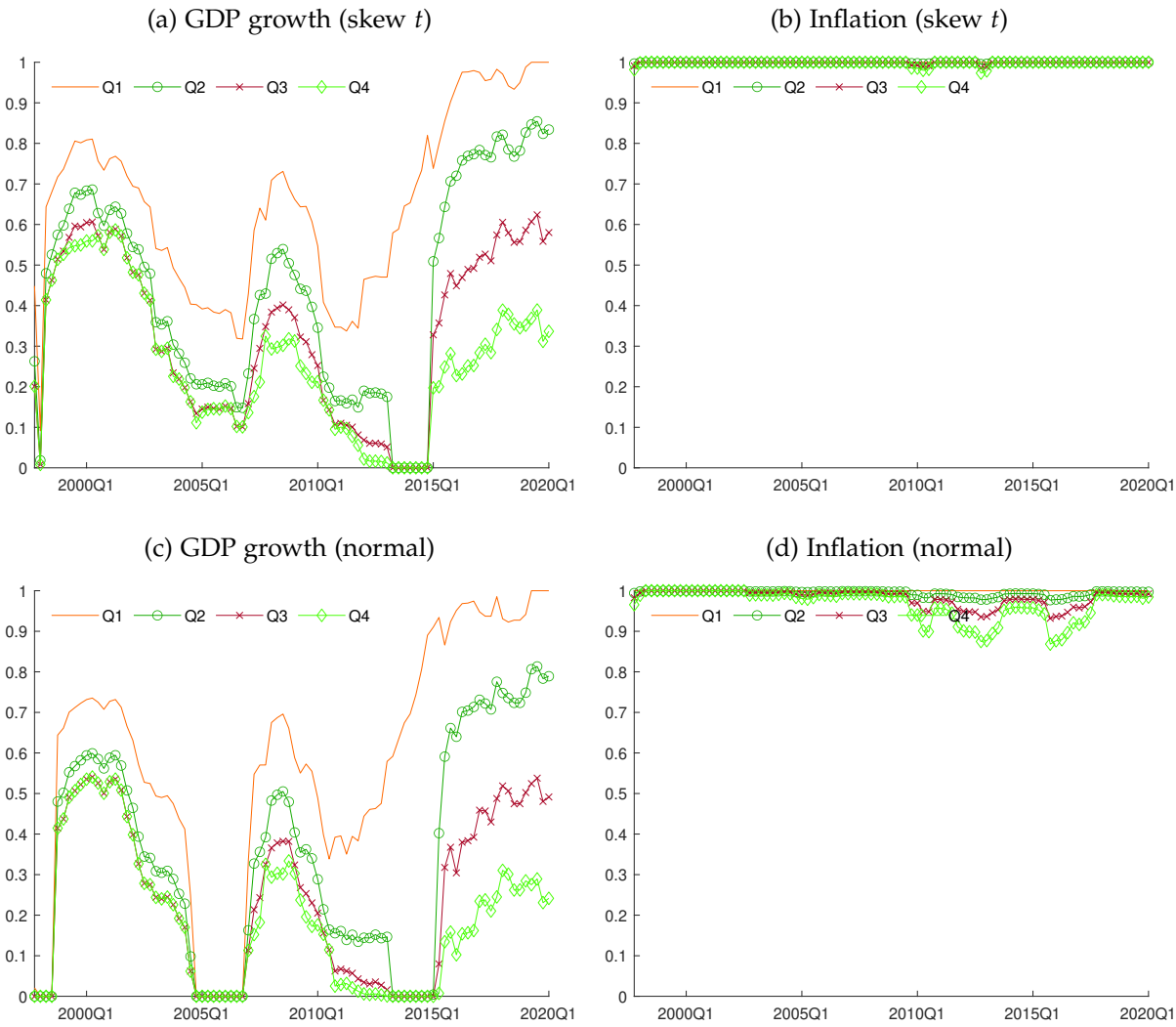
by

Gergely Ganics, Barbara Rossi, and Tatevik Sekhposyan

Appendix A Additional Results

Recall that in each SPF round, Equations (6) and (7) provide weight estimates $\widehat{w}^4_{q,0}$ for $q = 1, \ldots, 4$, displayed in Figure A.1.

Figure A.1: Weights on current year's density forecast

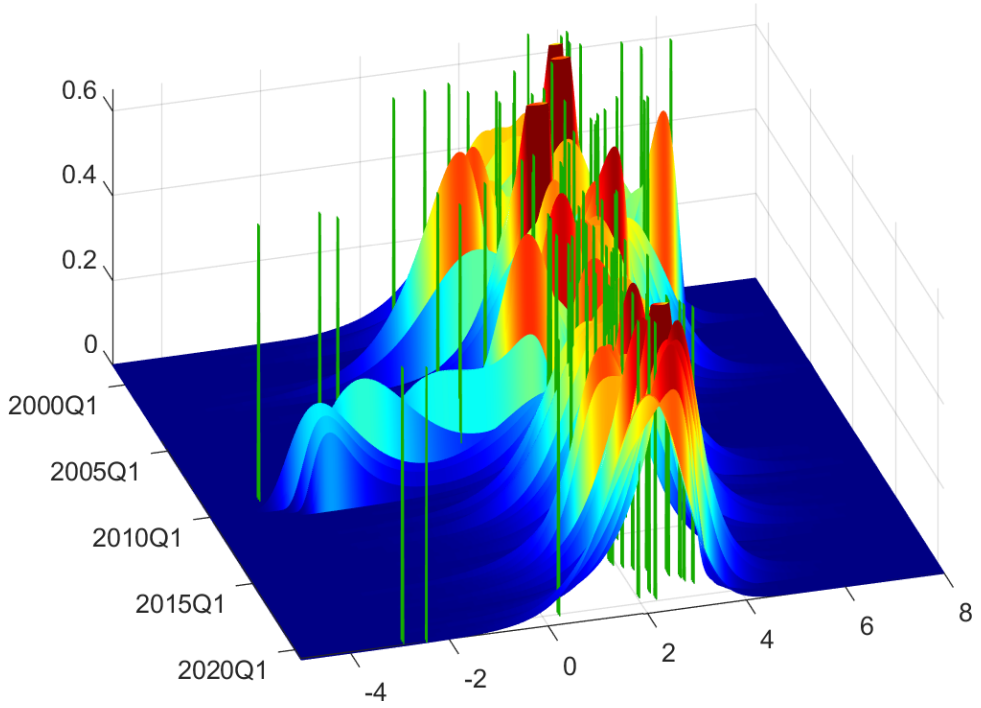


Note: The four panels in the figure depict the estimated combination weights on current year's density forecast corresponding to every quarter for each variable. Qj denotes the jth quarter in the year.

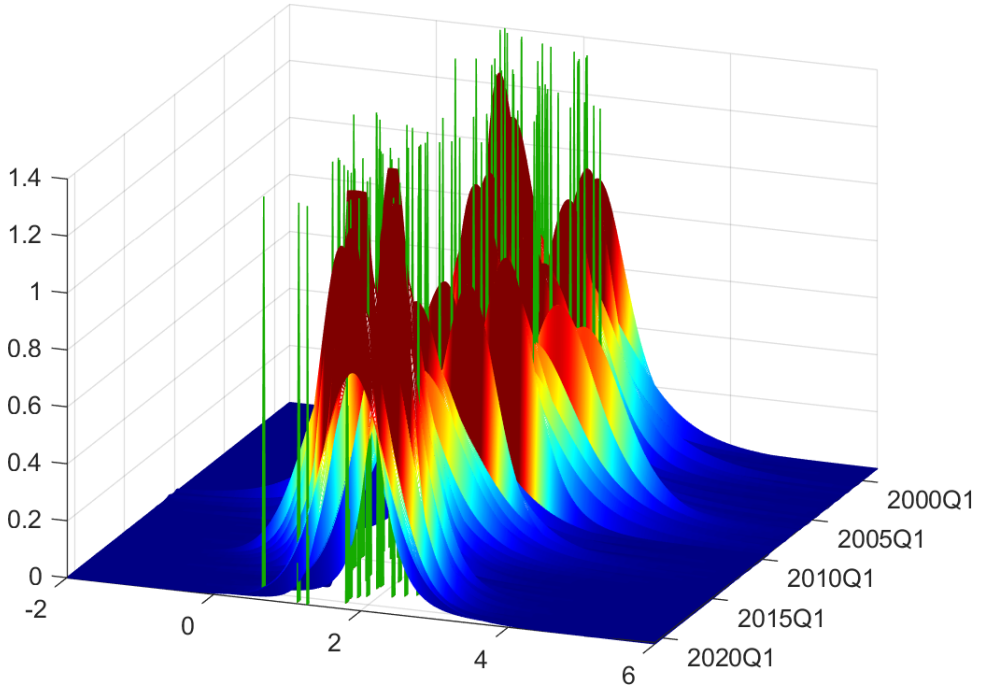
Figure A.2 displays the fixed-horizon densities for GDP growth and inflation over time. In Panel a, we can clearly see that during the recent Great Recession, the mean of the predictive distribution of GDP growth decreased and its dispersion increased. Furthermore, the densities are strikingly skewed 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 A.2: 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 2020:Q4.

Figure A.3 displays the various quantiles of the combined densities using the deterministic weights. In the case of inflation, comparing Panels b and d of Figure A.3 relative to Panels b and d of Figure 6 reveals that when the densities are combined based on the fixed weights, the quantiles are smoother and the combined density is more dispersed. For GDP growth, however, the density based on deterministic weights is much tighter, thus this combination, overall, understates the uncertainty relative to the combined density with estimated weights.

Figure A.4 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., 2020) 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 6) than the CMM model's before the Great Recession, but this reverses after the crisis.[20]

Figure A.7 displays the empirical CDF of the PITs along with the 95% confidence intervals, visually demonstrating the findings of Section 4.1 for the density combination schemes.

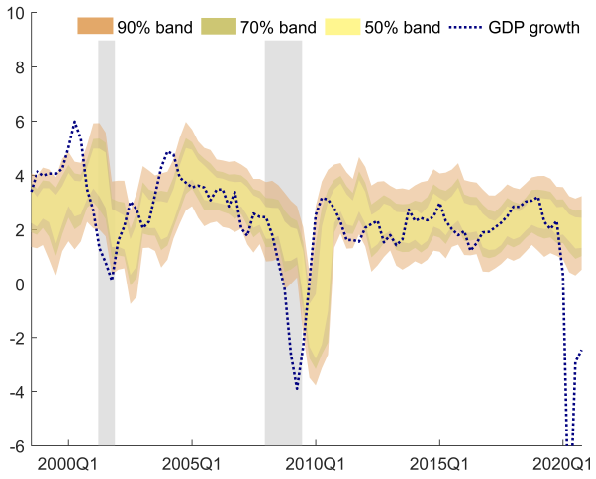
We have also evaluated the out-of-sample performance of the models using the quantile-weighted CRPS measure proposed by Gneiting and Ranjan (2011).[21] As Table A.1 shows, focusing attention on specific parts of the predictive distribution leads to similar results. This demonstrates that equal or superior predictive ability is not concentrated in specific regions of the predictive distributions, but rather it appears across the whole support.

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

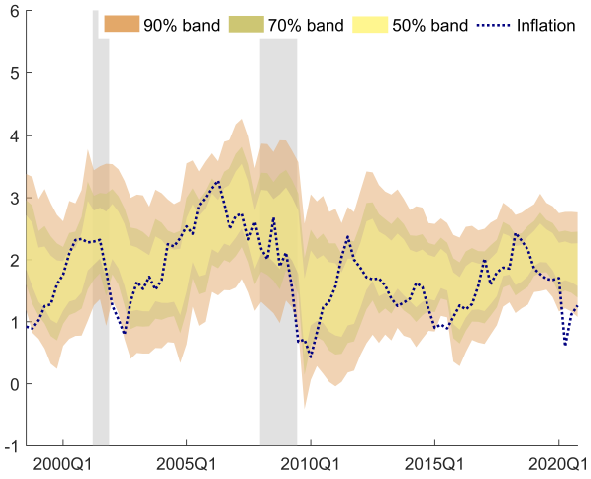
[21] We thank an anonymous referee for the suggestion.

Figure A.3: Predictive intervals of four-quarter-ahead combined predictive densities, using deterministic weights

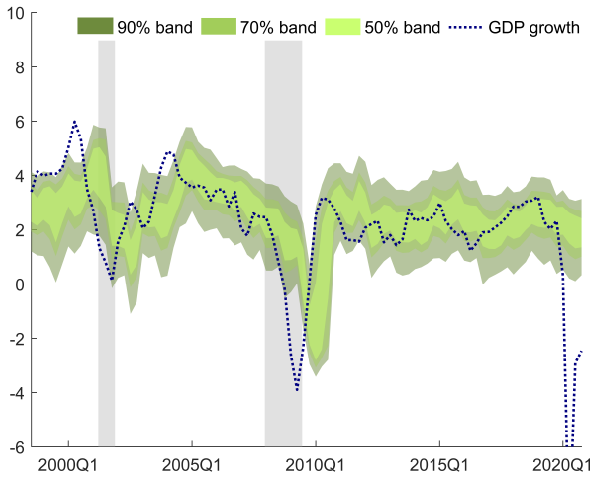
(a) GDP growth, normals



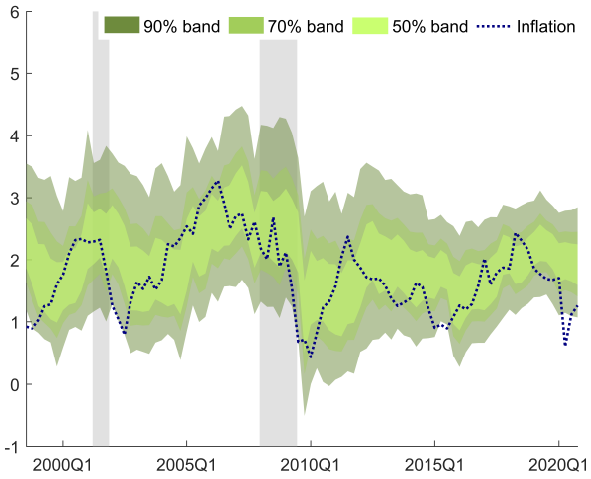
(b) Inflation, normals



(c) GDP growth, skew t



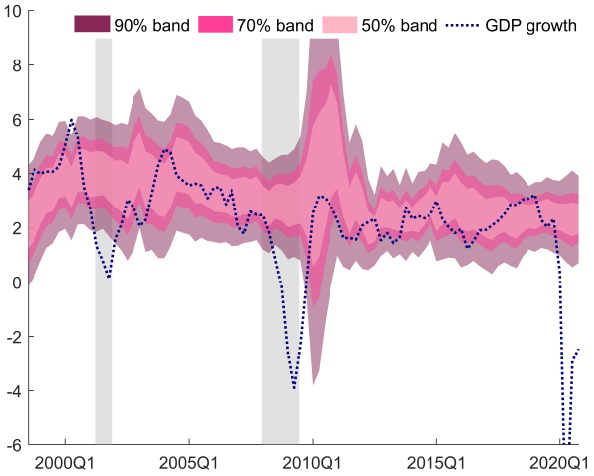
(d) Inflation, skew t



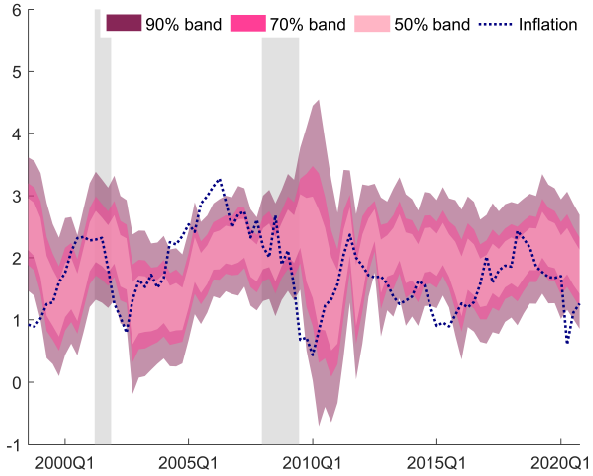
Note: The figure shows 90%, 70%, and 50% bands, corresponding to the 90%, 70%, and 50% equal-tailed predictive intervals of the combined four-quarter-ahead predictive densities for GDP growth (left column) and inflation (right column) based on the US SPF, using deterministic weights. The dotted (blue) 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 2020:Q4. Shaded areas denote NBER recession periods.

Figure A.4: Predictive intervals of four-quarter-ahead predictive densities of BVAR, PFE and CMM models

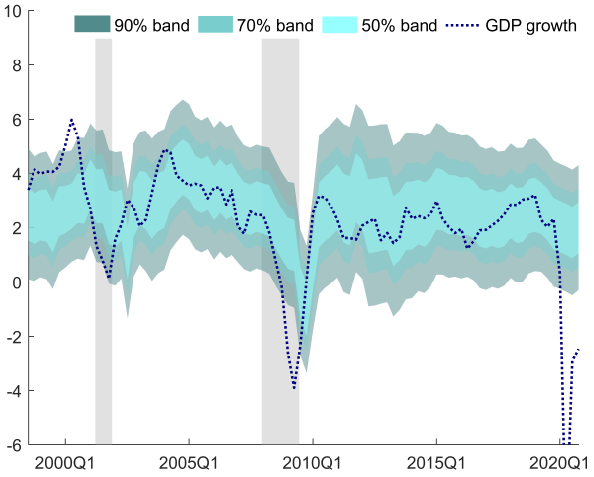
(a) GDP growth, BVAR



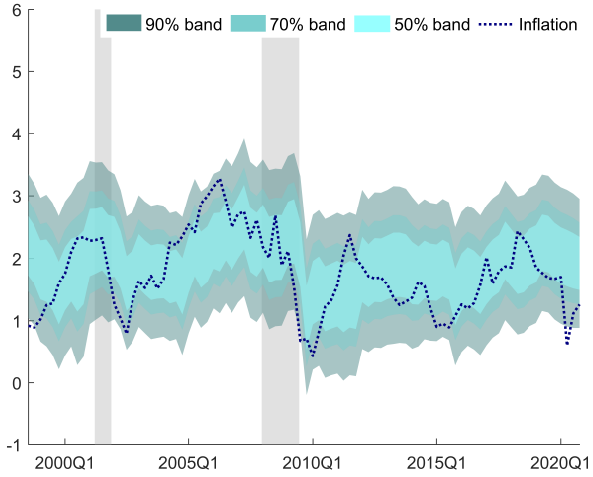
(b) Inflation, BVAR



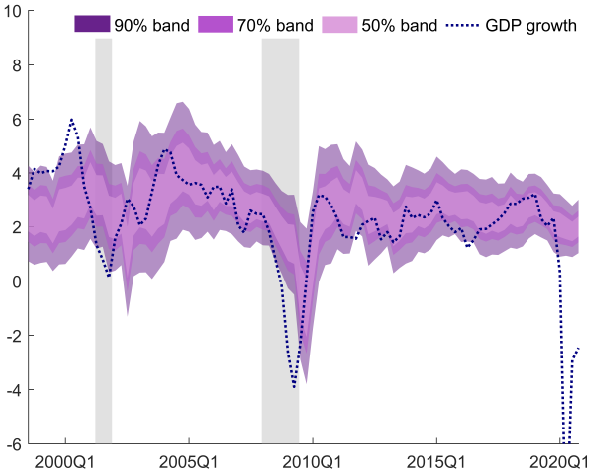
(c) GDP growth, PFE



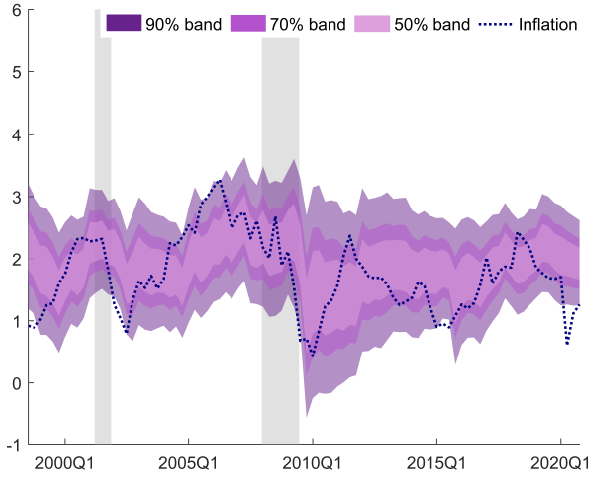
(d) Inflation, PFE



(e) GDP growth, CMM

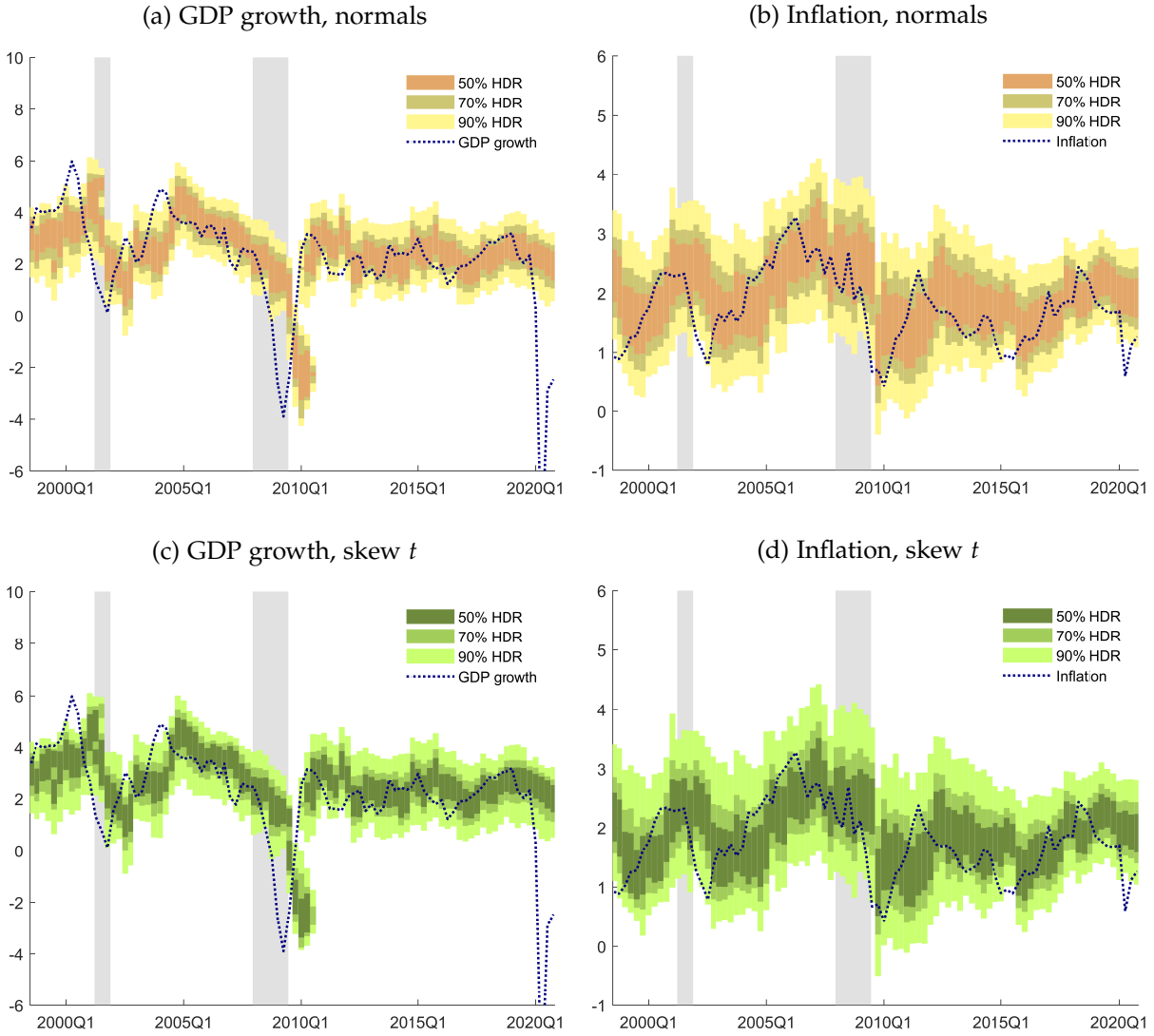


(f) Inflation, CMM



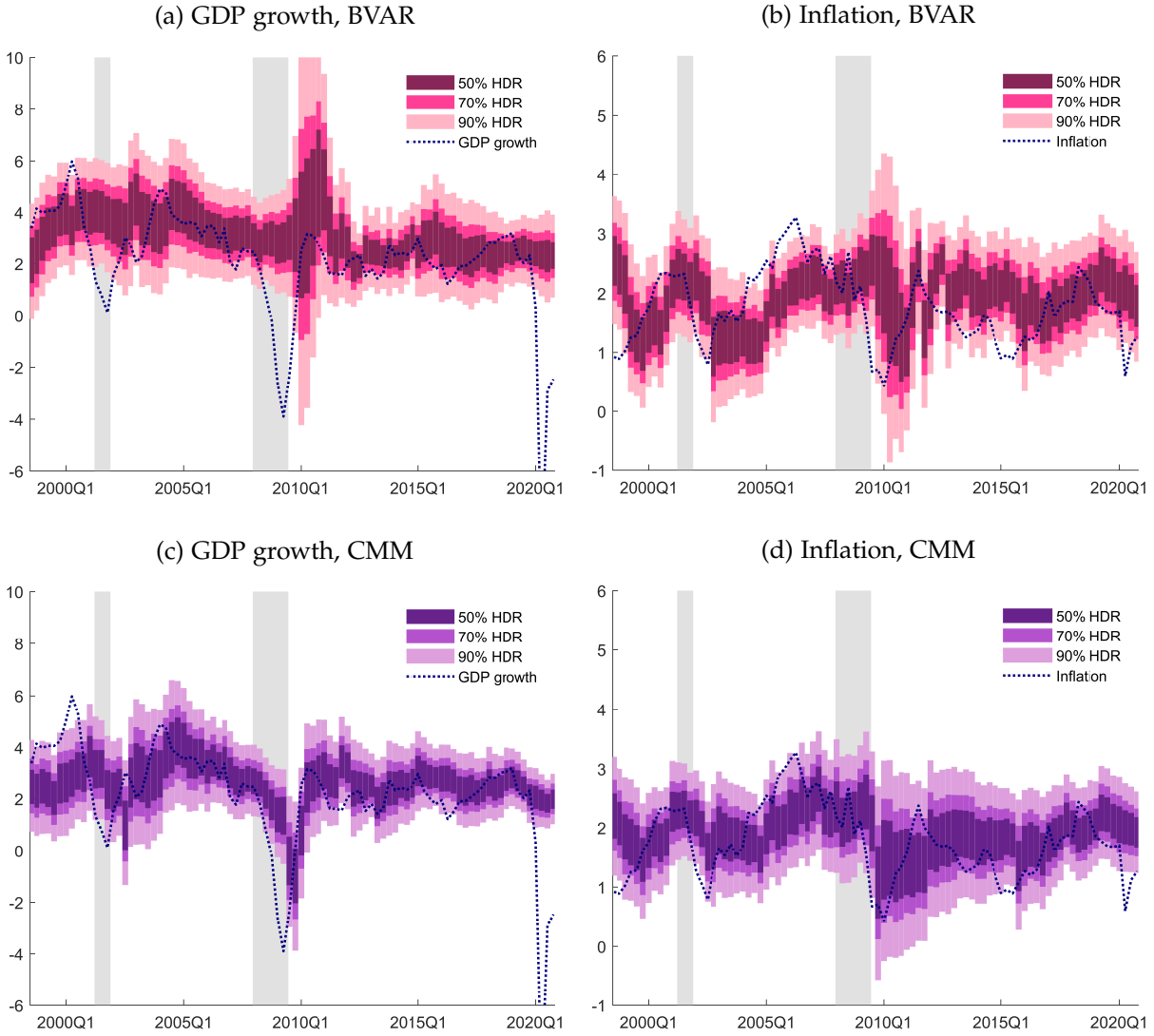
Note: The figure shows 90%, 70%, 50% bands, corresponding to the Bayesian VAR's (or the PFE model's or the CMM model's) 90%, 70% and 50% equal-tailed predictive intervals for GDP growth and inflation. The dotted blue lines mark the realized values of the variable of interest according to the first release. The forecast target dates on the horizontal axis range from 1998:Q3 to 2020:Q4. Shaded areas denote NBER recession periods.

Figure A.5: Highest density regions of four-quarter-ahead combined predictive densities, using deterministic 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 2020:Q4. Shaded areas denote NBER recession periods.

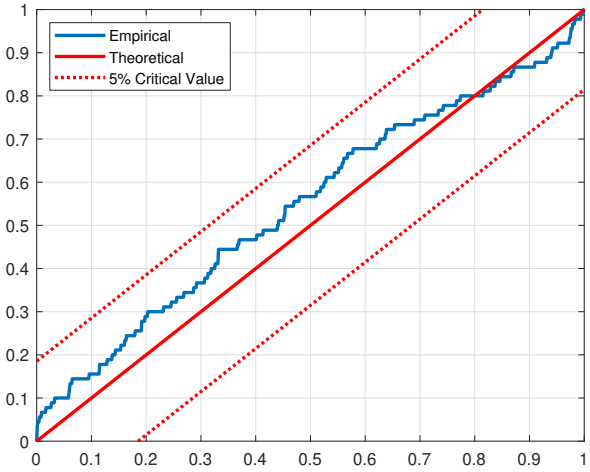
Figure A.6: Highest density regions of four-quarter-ahead predictive densities of BVAR and CMM models



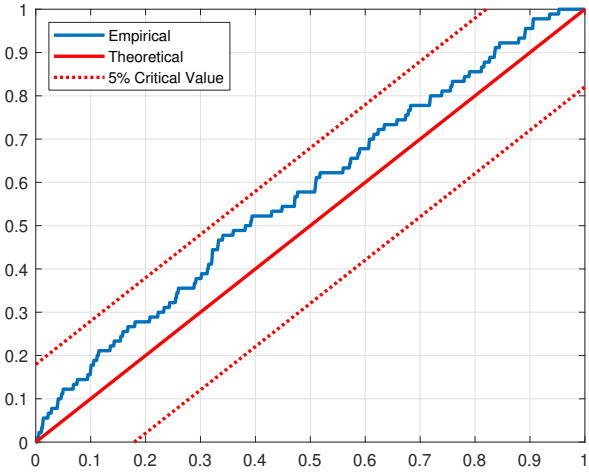
Note: The figure shows 90%, 70%, 50% highest density regions for GDP growth and inflation. The dotted blue lines mark the realized values of the variable of interest. The forecast target dates on the horizontal axis range from 1998:Q3 to 2020:Q4. Shaded areas denote NBER recession periods.

Figure A.7: Uniformity of the PITs of the combined skew t distributions

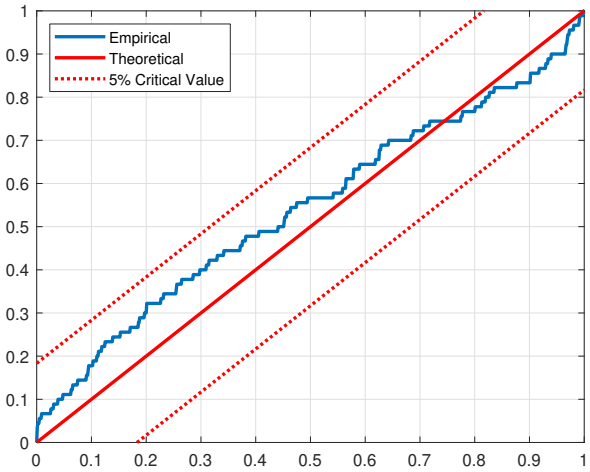
(a) GDP growth, skew t, estimated weights



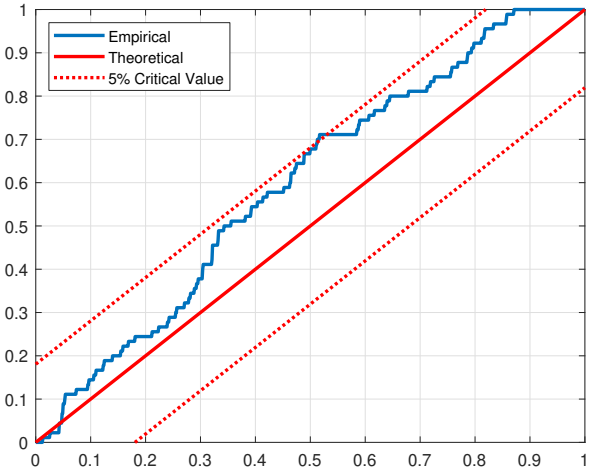
(b) Inflation, skew t, estimated weights



(c) GDP growth, skew t, deterministic weights



(d) Inflation, skew t, deterministic weights



Note: The figure shows the empirical CDFs of PITs of the combined skew t distributions for GDP growth and inflation based on our weight estimation scheme (upper panels) and the deterministic weighting procedure (bottom panels), the CDF of the PITs under the null hypothesis of correct calibration (the 45 degree line) and the 5% critical values bands (dotted lines) based on the Kolmogorov-Smirnov test in Rossi and Sekhposyan (2019), using their bootstrapped critical values.

Table A.1: Relative forecast evaluation: CRPS

	GDP growth				
	full	central	tails	left	right
N vs N (det)	−0.96(0.17)	−1.08(0.14)	−0.34(0.37)	−0.79(0.21)	−0.93(0.18)
ST vs ST (det)	−1.24(0.11)	−1.40(0.08)*	−0.40(0.34)	−1.18(0.12)	−0.90(0.18)
N vs BVAR	−2.07(0.02)**	−2.07(0.02)**	−1.98(0.02)**	−1.82(0.03)**	−1.98(0.02)**
ST vs BVAR	−2.19(0.01)**	−2.17(0.02)**	−2.10(0.02)**	−2.00(0.02)**	−2.00(0.02)**
N vs PFE	0.16(0.56)	0.07(0.53)	0.23(0.59)	−2.22(0.01)**	1.46(0.93)
ST vs PFE	−0.06(0.48)	−0.05(0.48)	−0.06(0.48)	−2.29(0.01)**	1.35(0.91)
N vs CMM	0.91(0.82)	0.87(0.81)	0.93(0.82)	0.52(0.70)	1.04(0.85)
ST vs CMM	0.58(0.72)	0.63(0.74)	0.48(0.68)	0.18(0.57)	0.80(0.79)
	Inflation				
	full	central	tails	left	right
N vs N (det)	1.24(0.89)	1.12(0.87)	1.28(0.90)	−0.36(0.36)	1.99(0.98)
ST vs ST (det)	1.46(0.93)	1.49(0.93)	1.14(0.87)	−0.28(0.39)	2.30(0.99)
N vs BVAR	−3.01(0.00)***	−3.13(0.00)***	−2.47(0.01)***	−2.74(0.00)***	−2.60(0.00)***
ST vs BVAR	−2.92(0.00)***	−2.98(0.00)***	−2.55(0.01)***	−2.59(0.00)***	−2.67(0.00)***
N vs PFE	0.92(0.82)	0.77(0.78)	1.19(0.88)	−0.27(0.39)	1.62(0.95)
ST vs PFE	1.14(0.87)	1.02(0.85)	1.32(0.91)	0.06(0.52)	1.74(0.96)
N vs CMM	1.00(0.84)	0.82(0.80)	1.43(0.92)	0.34(0.63)	1.55(0.94)
ST vs CMM	1.21(0.89)	1.06(0.86)	1.58(0.94)	0.63(0.74)	1.66(0.95)

Note: The top panel displays displays the Diebold and Mariano (1995) and West (1996) test statistics and p-values (in parentheses, with rejection region in the left tail) comparing predictive accuracy measured by Gneiting and Ranjan's (2011) weighted CRPS for GDP growth, and the bottom panel for inflation. In the headers, "full" refers to the standard CRPS, "central", "tails", "left" and "right" emphasize attention on specific regions of interest. For an explanation of the different abbreviations, please see the main text. 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 2020:Q1, with corresponding realizations between 1998:Q3 and 2020:Q4.

Appendix B Robustness Checks

B.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

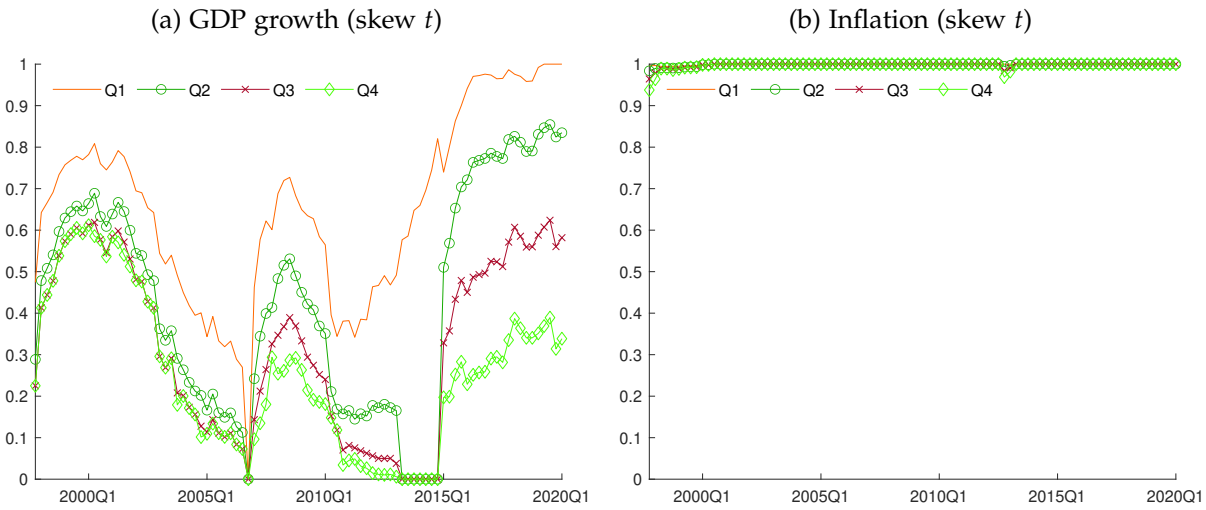
$$f(x;\mu,\sigma,\alpha,\nu) = \frac{2}{\sigma} t_\nu\left(\frac{x-\mu}{\sigma}\right) T_{\nu+1}\left(\alpha\frac{x-\mu}{\sigma}\sqrt{\frac{\nu+1}{\nu+\left(\frac{x-\mu}{\sigma}\right)^2}}\right), \tag{19}$$

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 B.1 are very similar to their counterparts in Figure A.1.

Figure B.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.

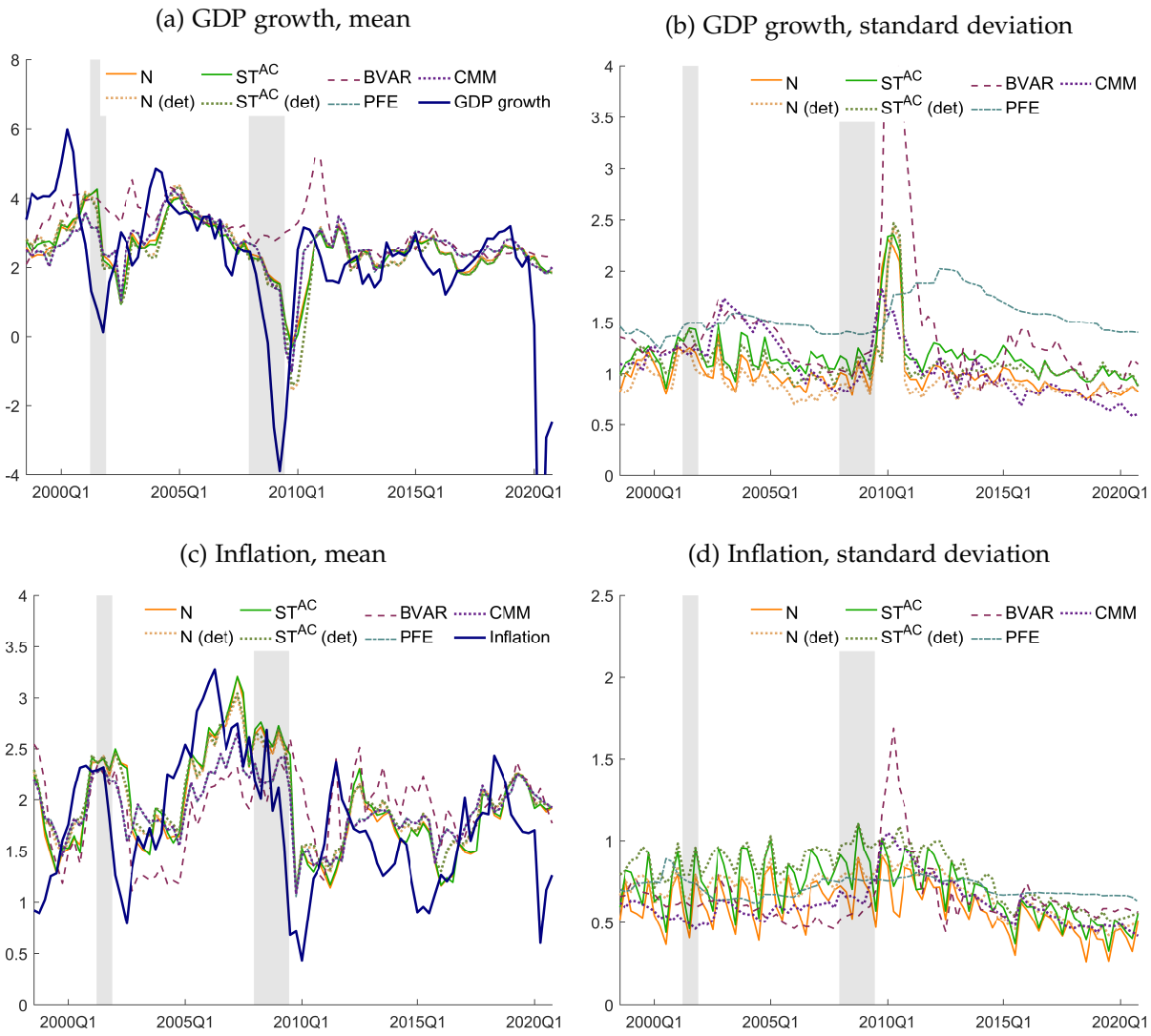
Figure B.2 shows that when using Azzalini and Capitanio's (2003) skew t distribution, the means and standard deviations of the mixture distribution are very similar to the case when using Jones and Faddy's (2003) skew t distribution in the main text, both for GDP growth and inflation.

Figures B.3 and B.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 B.5.

Figure B.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.

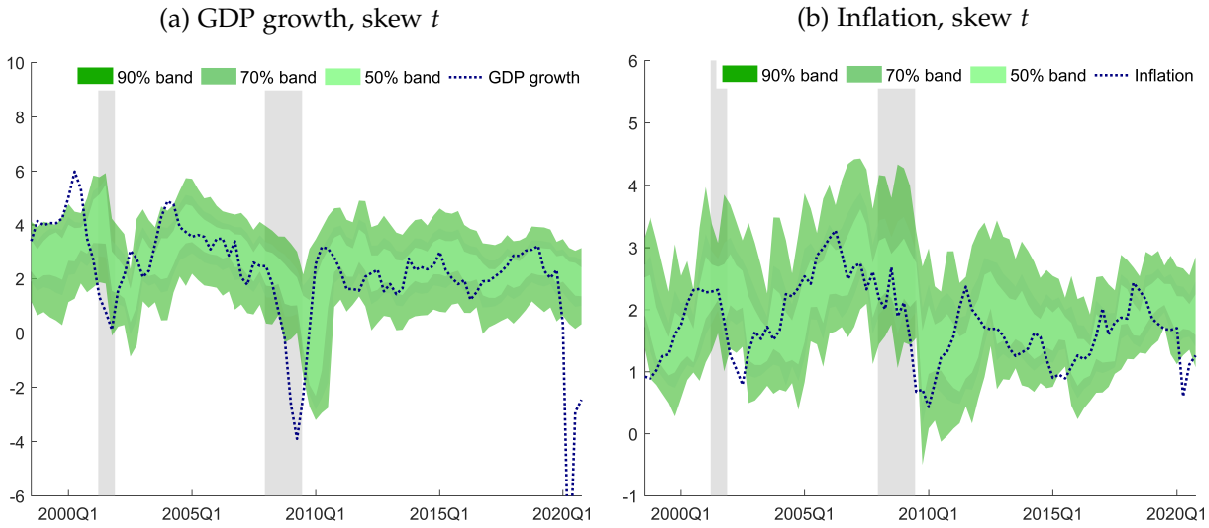
Tables B.1 to B.4 display the same forecast evaluation statistics as in the main text, adding the results with the Azzalini and Capitanio (2003) distribution. Furthermore, we show results for the BVAR and CMM models, where the parameters are estimated based on a recursive estimation scheme. As we can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

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



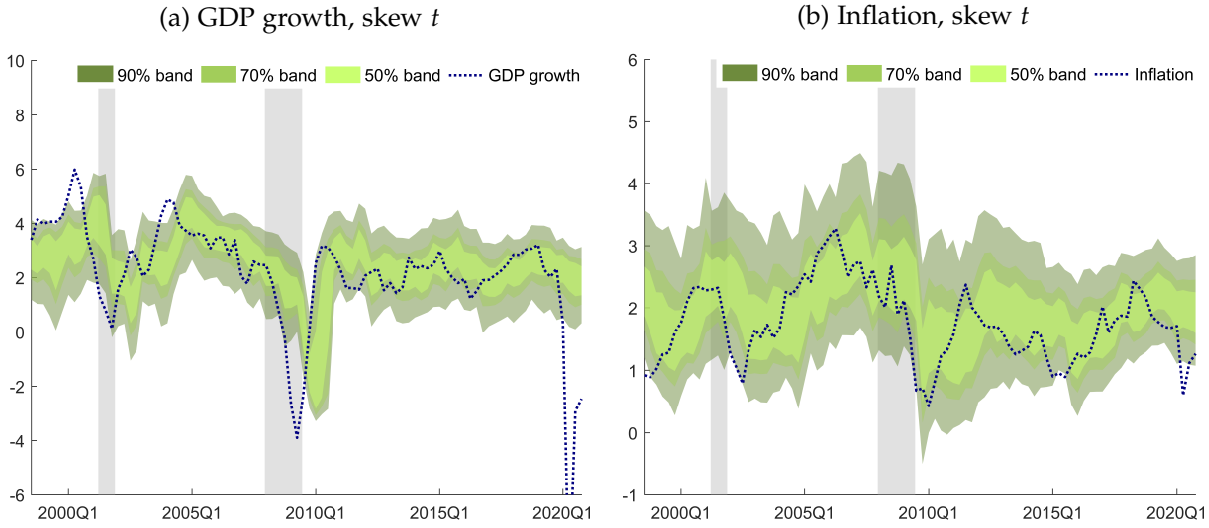
Note: The figures show the mean and the standard deviation of the four-quarter-ahead GDP growth forecasts (subfigures a and b) and inflation forecasts (subfigures c and d) of various methods at the corresponding *target* dates ranging from 1998:Q3 to 2020:Q4. For an explanation of the different abbreviations, please see the main text. Shaded areas are NBER recession periods.

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



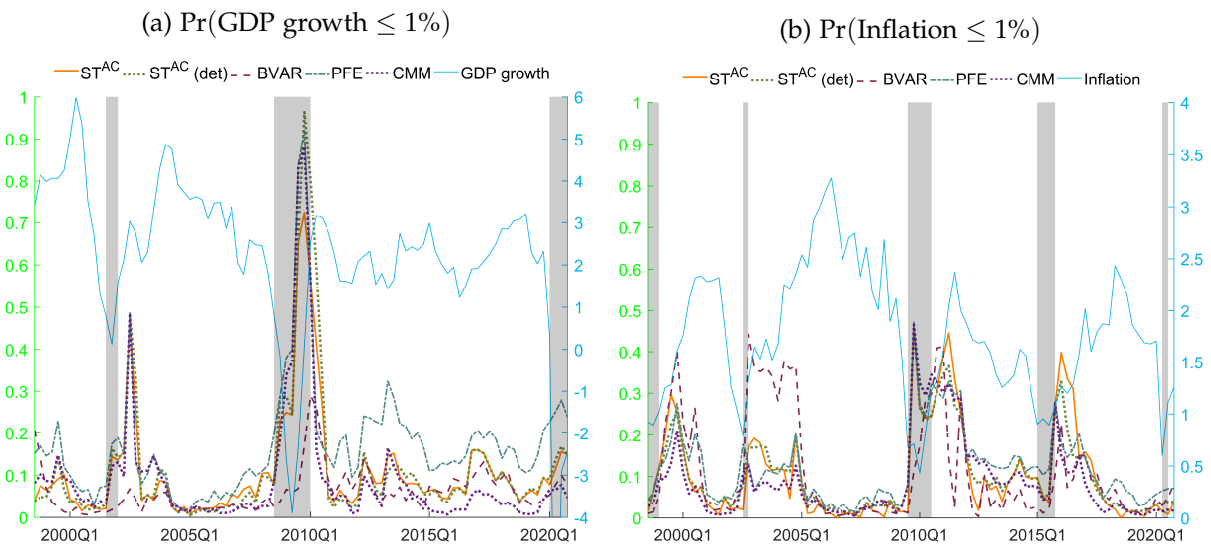
Note: The figure shows 90%, 70%, and 50% bands, corresponding to the 90%, 70%, and 50% equal-tailed predictive intervals of the combined four-quarter-ahead predictive densities for GDP growth (left column) and inflation (right column) based on the US SPF, using the proposed weight estimator. The dotted (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 2020:Q4. Shaded areas denote NBER recession periods.

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



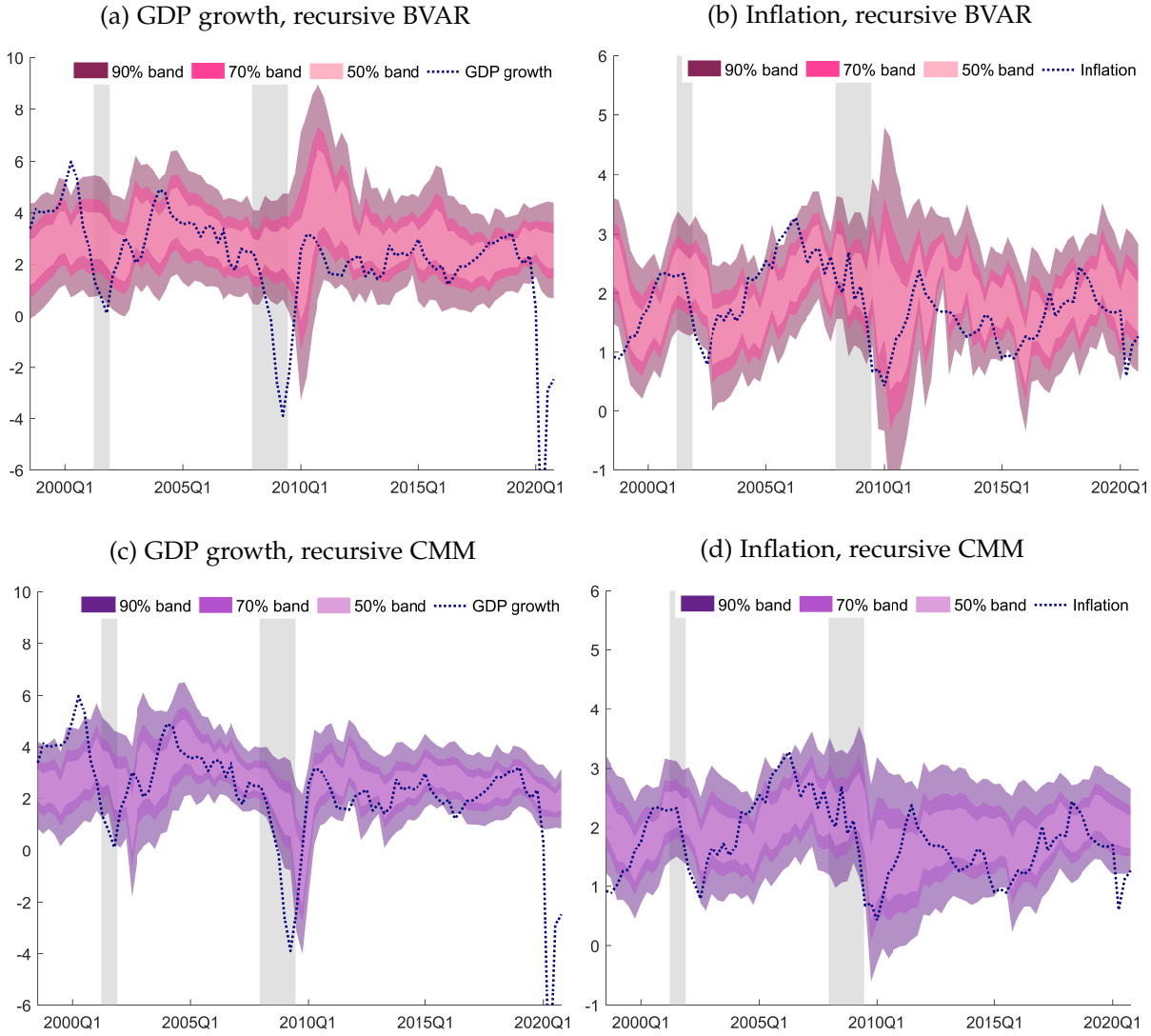
Note: The figure shows 90%, 70%, and 50% bands, corresponding to the 90%, 70%, and 50% equal-tailed predictive intervals of the combined four-quarter-ahead predictive densities for GDP growth (left column) and inflation (right column) based on the US SPF, using deterministic weights. The dotted (blue) 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 2020:Q4. Shaded areas denote NBER recession periods.

Figure B.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 2020:Q4. Shaded grey areas denote the periods when the predicted event (e.g. GDP growth $\leq$ 1%) did in fact occur.

Figure B.6: Predictive intervals of four-quarter-ahead predictive densities of recursively estimated BVAR and 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 2020:Q4. Shaded areas are NBER recession periods.

Table B.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 (det)	**0.96**(**0.47**)	**0.26**(**0.46**)	1.68(0.06)	0.90(0.05)
ST^{JF}(det)	**1.03**(**0.40**)	**0.29**(**0.42**)	1.71(0.05)	0.82(0.06)
ST^{AC}(det)	**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 p-values of Kolmogorov–Smirnov (KS) and Cramér–von Mises (CvM) tests of the null hypothesis of uniformity of PITs for different target variables (in the column headers) and models (in rows). N and ST correspond to the combinations of normal and skew t distributions using our proposed weight estimates (see Section 3.2), while N (det) and ST (det) denote their conterparts using deterministic weights (see Section 3.3.1). PFE corresponds to the normal distribution based on past forecast errors (Section 3.3.3), BVAR is the Bayesian VAR of Section 3.3.2, while CMM is the stochastic volatility model based on point forecast revisions (Section 3.3.4). 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 2020:Q1, with corresponding realizations between 1998:Q3 and 2020:Q4.

Table B.2: Absolute forecast evaluation: coverage

	GDP growth			Inflation		
	50%	70%	90%	50%	70%	90%
N	**49.4**(**0.92**)	**67.1**(**0.66**)	78.5(0.04)	**55.7**(**0.36**)	**73.4**(**0.54**)	**88.6**(**0.73**)
ST^{JF}	**48.1**(**0.77**)	**63.3**(**0.31**)	**83.5**(**0.18**)	**53.2**(**0.61**)	**73.4**(**0.52**)	**88.6**(**0.74**)
ST^{AC}	**48.1**(**0.77**)	**63.3**(**0.31**)	**82.3**(**0.13**)	**53.2**(**0.61**)	**73.4**(**0.52**)	**89.9**(**0.98**)
N (det)	**41.8**(**0.19**)	57.0(0.06)	78.5(0.03)	63.3(0.04)	81.0(0.03)	**92.4**(**0.48**)
ST^{JF}(det)	**41.8**(**0.19**)	57.0(0.06)	81.0(0.08)	60.8(0.09)	81.0(0.03)	**91.1**(**0.75**)
ST^{AC}(det)	**41.8**(**0.19**)	57.0(0.06)	79.8(0.05)	60.8(0.09)	81.0(0.03)	**92.4**(**0.48**)
BVAR (rolling)	**46.8**(**0.62**)	**69.6**(**0.95**)	**88.6**(**0.76**)	**40.5**(**0.14**)	55.7(0.03)	76.0(0.01)
BVAR (recursive)	**49.4**(**0.93**)	**67.1**(**0.66**)	**88.6**(**0.76**)	**51.9**(**0.78**)	**68.4**(**0.78**)	**83.5**(**0.15**)
PFE	65.8(0.02)	**77.2**(**0.22**)	**93.7**(**0.29**)	63.3(0.04)	**77.2**(**0.20**)	96.2(0.01)
CMM (rolling)	**49.4**(**0.93**)	**59.5**(**0.12**)	**84.8**(**0.26**)	**51.9**(**0.77**)	**69.6**(**0.95**)	**87.3**(**0.54**)
CMM (recursive)	**48.1**(**0.78**)	**59.5**(**0.11**)	**84.8**(**0.26**)	**51.9**(**0.77**)	**73.4**(**0.56**)	**89.9**(**0.97**)

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 the definition of the model abbreviations, see Table 1. 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 2020:Q1, with corresponding realizations between 1998:Q3 and 2020:Q4.

Table B.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 (det)	0.79	0.33
ST^{JF}(det)	0.79	0.33
ST^{AC}(det)	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 (det)	$-0.99(0.16)$	$0.98(0.84)$
ST^{JF} vs ST^{JF}(det)	$-1.35(0.09)^{*}$	$1.19(0.88)$
ST^{AC} vs ST^{AC}(det)	$-1.37(0.09)^{*}$	$1.08(0.86)$
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 the definition of the model abbreviations, see Table 1. The bottom panel displays the Diebold and Mariano (1995) and West (1996) 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 2020:Q1, with corresponding realizations between 1998:Q3 and 2020:Q4.

Table B.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 (det)	0.076	0.115
ST^{JF}(det)	0.076	0.113
ST^{AC}(det)	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 (det)	$-0.59(0.28)$	$2.63(1.00)$
ST^{JF} vs ST^{JF}(det)	$-0.55(0.29)$	$2.70(1.00)$
ST^{AC} vs ST^{AC}(det)	$-0.64(0.26)$	$2.67(1.00)$
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)	$2.26(0.99)$	$2.91(1.00)$

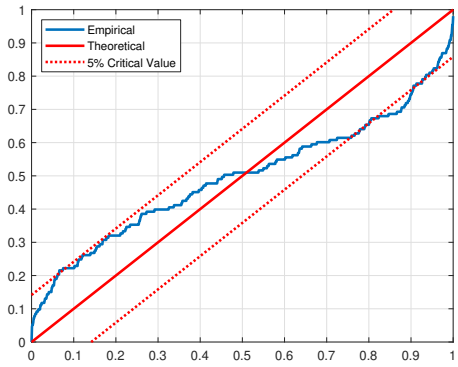
Note: The target variable used for both estimation and forecast evaluation and the corresponding extreme event are shown in the column headers. The top panel displays the Brier score of various density combination methods in the rows. For each variable, the lowest value is in bold. For the definition of the model abbreviations, see Table 1. The bottom panel displays the Diebold and Mariano (1995) and West (1996) 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 2020:Q1, with corresponding realizations between 1998:Q3 and 2020:Q4.

Appendix C Further Robustness Results

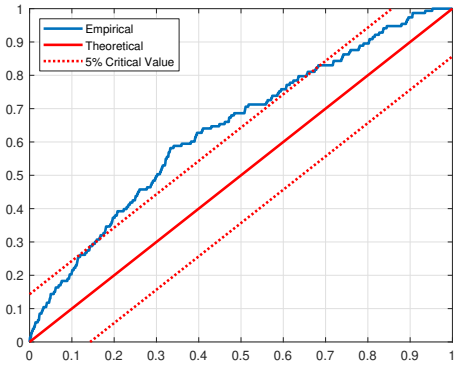
This section shows further robustness studies, where we assess the correct calibration of the component as well as combined densities in Figure C.1, show an example of local identification of the combination weights in Figure C.2, and in display the weight estimates when using $R = 80$ observations instead of $R = 60$ observations in each rolling window.

Figure C.1: Calibration of component distributions

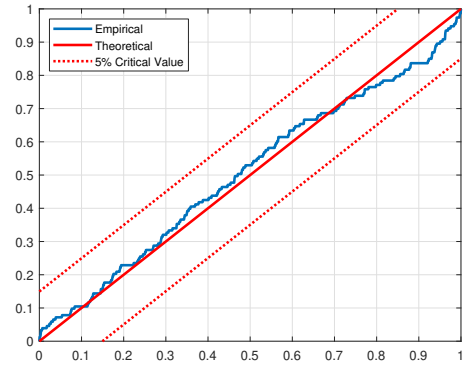
(a) GDP growth, current year



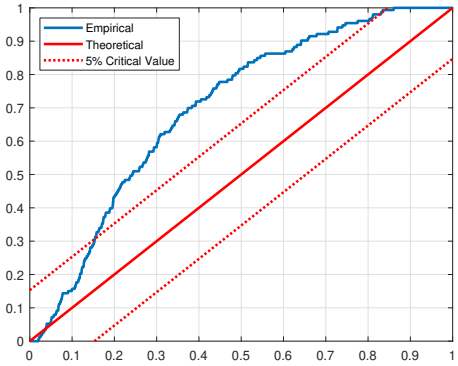
(b) Inflation, current year



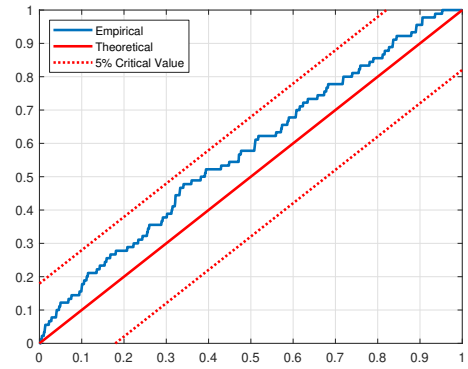
(c) GDP, next year



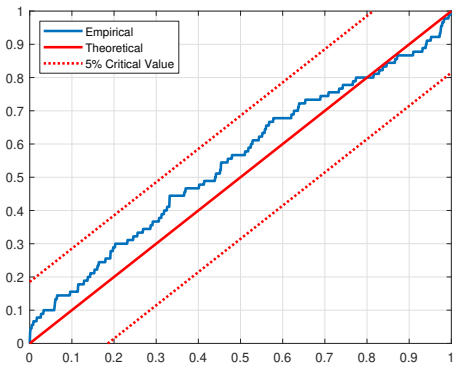
(d) Inflation, next year



(e) GDP growth, four-quarter-ahead

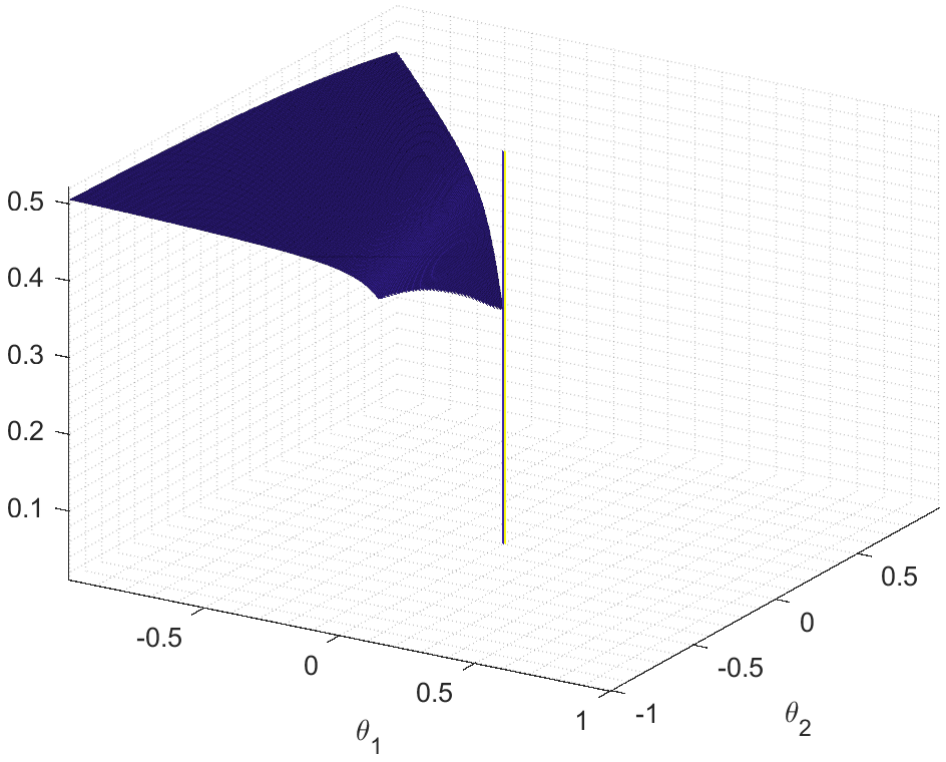


(f) Inflation, four-quarter-ahead



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. The last row depicts similar results for four-quarter-ahead density combinations, where components are based on a skew t approximation and combined with optimal weights.

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

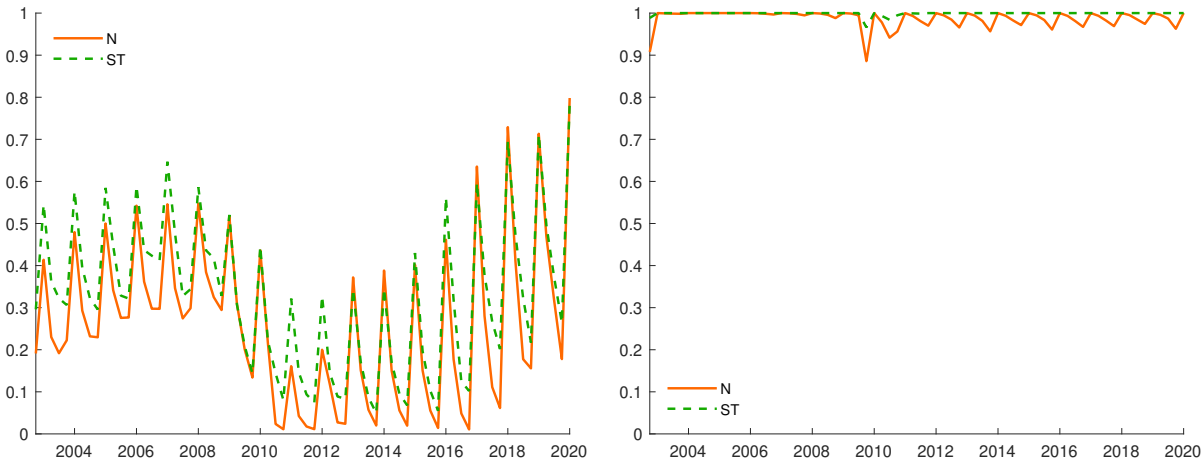


Note: The figure shows the value of the objective function in Equation (7) (vertical axis) as a function of the parameter vector (θ_1, θ_2) when combining Jones and Faddy (2003) skew t distributions of inflation in the 1997:Q4 SPF round. The vertical spike at $(0.0015, -0.0015)$ marks the optimum.

Figure C.3: Weights on current year's density forecast using, rolling window size $R = 80$

(a) GDP growth (normal and skew t)

(b) Inflation (normal and skew t)



Note: The panels in the figure depict the estimated combination weights one would apply in each SPF round (horizontal axis) to current year's density to combine the forecasts.

Appendix D An Application to BVAR Forecasts

To investigate the properties of our proposed density combination scheme, we performed the following exercise. We estimated the BVAR model in a rolling window scheme, aligning the BVAR's information set with that of the SPF panelists between the 1981:Q3 and 2020:Q1 SPF rounds. Then we constructed the BVAR's fixed-event forecasts for GDP growth and inflation for the current and the next calendar year, mimicking the SPF questionnaire, to obtain annual average over annual average growth rate forecasts, and fitted a normal distribution to the resulting MCMC draws. Next, we applied our proposed weight estimation scheme on these fitted distributions, in exactly the same manner as we did in the case of the SPF forecasts (see Section 2).

Figure D.1 shows the results of Rossi and Sekhposyan's (2019) test on the uniformity of the PITs. As we can see, next year's (fixed-event) forecasts are better calibrated, and (our combination method) delivers correctly calibrated fixed-horizon density forecasts by putting more weight on next year's forecasts relative to this year's forecasts.

As Table D.1 demonstrates, our combination method (third row) can indeed improve upon the individual components' calibration (first two rows), but it certainly has limitations, and if the component densities are considerably miscalibrated, then the optimal weighting scheme cannot fully correct for this fact. It should be noted that the tests in this table are carried out at 10%, slightly more liberal than the 5% significance level used in Figure D.1.

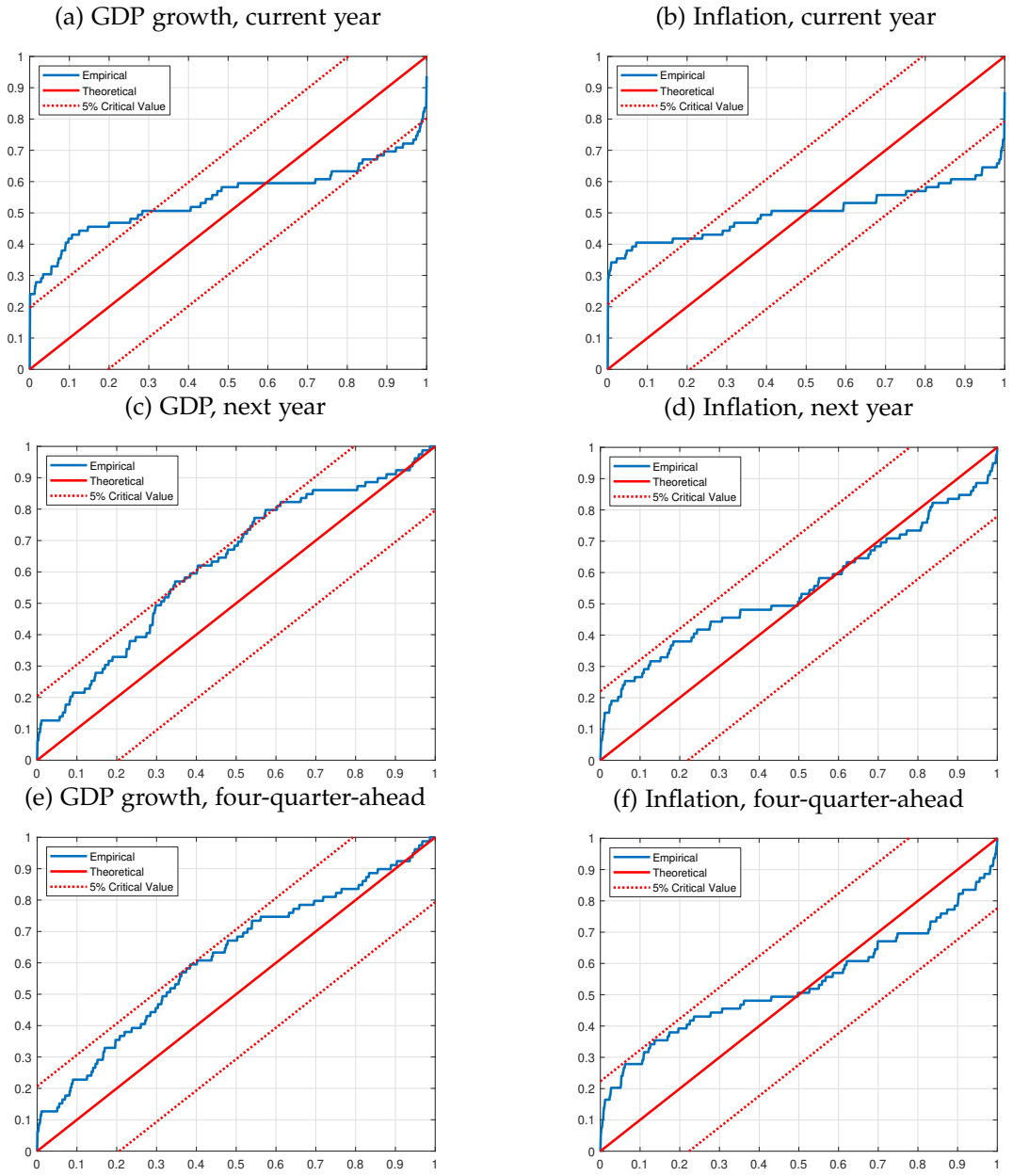
Table D.1: Absolute forecast evaluation: uniformity of PIT

	GDP growth		Inflation	
	KS	CvM	KS	CvM
BVAR (current year)	2.87(0.00)	2.80(0.01)	2.97(0.00)	3.42(0.02)
BVAR (next year)	2.01(0.03)	1.65(0.02)	1.74(0.10)	**0.84**(**0.21**)
BVAR (optimal weights)	1.87(0.04)	1.32(0.04)	1.93(0.06)	**1.12**(**0.15**)
BVAR (original)	2.29(0.00)	1.86(0.00)	**1.27**(**0.28**)	**0.28**(**0.50**)

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). 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 forecast origins range from 1997:Q4 to 2020:Q1, with corresponding realizations between 1998:Q3 and 2018:Q1.

Considering GDP growth, Figures D.2 and D.3 show the BVAR's fixed-event forecasts (Figures D.2a, D.2b, D.3a and D.3b), and the optimally weighted combination (solid blue line)

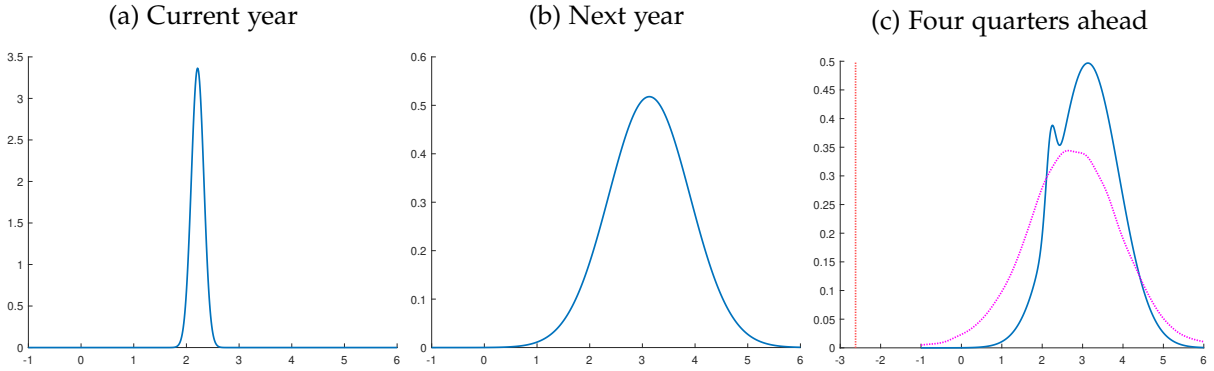
Figure D.1: Calibration of BVAR component distributions



Note: The figure shows the empirical CDFs of the current year and next year densities for GDP growth and inflation based on the BVAR (after fitting a normal 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. The last row depicts similar results for four-quarter-ahead density combinations, where components are combined with optimal weights.

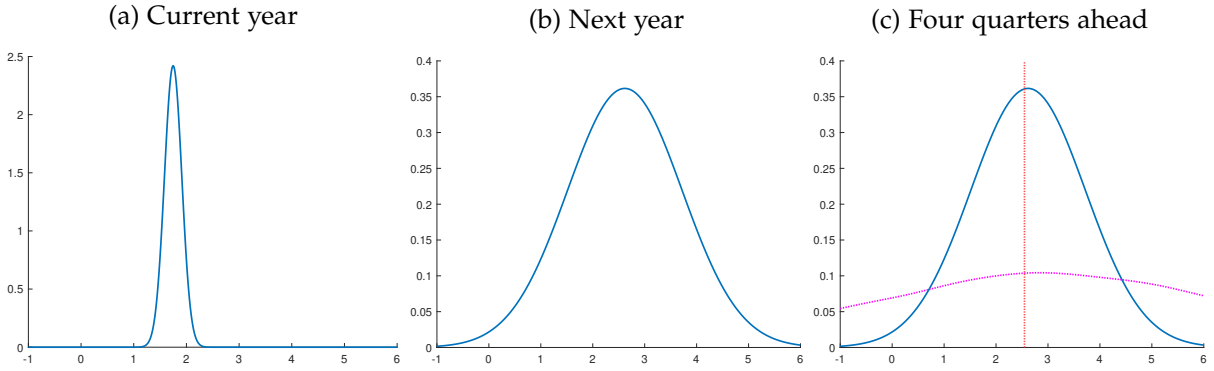
along with the BVAR's fixed-horizon forecasts (dashed magenta line) in Figures D.2c and D.3c. Essentially, the blue solid line is what our proposed procedure would predict if we were given annual average over annual average forecasts from the BVAR, while the dashed magenta line would be the outcome of the original BVAR that is forecasting four-quarter-ahead, quarter over quarter growth rates. As the picture shows, our mixing strategy can result in multi-modal (in this case bimodal) densities (Figure D.2c for GDP growth), even if the underlying model

Figure D.2: BVAR forecasts of GDP growth and their combination as of 2008:Q2



Note: The plots display the BVAR's fixed-event forecasts (panels (a) and (b)), along with their optimally weighted combination (solid blue line in panel (c)), and the BVAR's fixed-horizon forecast (dashed magenta line in panel (c)). The dashed vertical red line corresponds to the actual realization.

Figure D.3: BVAR forecasts of GDP growth and their combination as of 2009:Q2

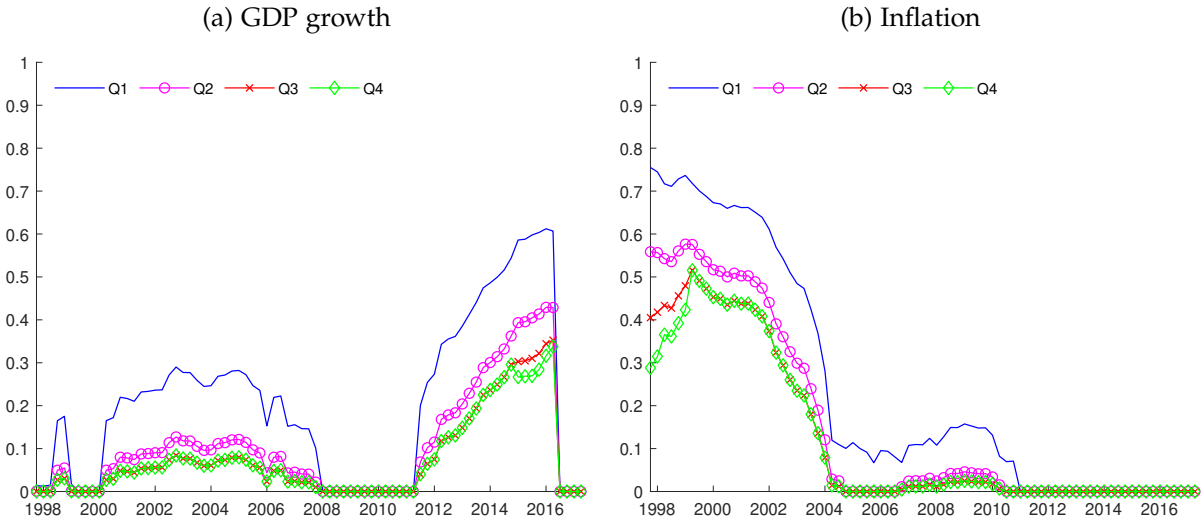


Note: The plots display the BVAR's fixed-event forecasts (panels (a) and (b)), along with their optimally weighted combination (solid blue line in panel (c)), and the BVAR's fixed-horizon forecast (dashed magenta line in panel (c)). The dashed vertical red line corresponds to the actual realization.

does not prescribe multi-modality to the fixed-horizon density forecasts. Part of the reason is that the current year and next year GDP growth forecasts, depicted in Figures D.2a and D.2b, respectively, appear to be dramatically different from each other. However, we argue that when we operate with the objective of obtaining correct calibration of the predictive densities, then allowing for multi-modality is a desired feature. Importantly, our objective is *not* obtaining the BVAR's underlying fixed-horizon predictive density, but rather a correctly calibrated one, which might easily differ from the BVAR's density. It is worth to emphasize that the distributions based on the original BVAR are not obtained based on correct calibration criterion either.

The estimated weights that are used to generate Figure D.2c and Figure D.3c are 0.04 and 0.96 for the current year and next year forecasts in 2008:Q2, and 0.00 and 1.00 in 2009:Q2. Figure D.4 shows the estimated weights over time for both GDP growth (Figure D.4a) and

Figure D.4: Estimated combination weights of BVAR forecasts



Note: The four lines 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.

inflation (Figure D.4b), to be interpreted analogously to Figure A.1.