Journal of Forecasting

WILEY

RESEARCH ARTICLE OPEN ACCESS

Regime-Switching Density Forecasts Using Economists' Scenarios

Graziano Moramarco

Department of Economics, University of Bologna, Bologna, Italy

Correspondence: Graziano Moramarco (graziano.moramarco@unibo.it)

Received: 15 September 2021 | **Revised:** 5 October 2024 | **Accepted:** 27 October 2024

Keywords: density forecasts | forecast combinations | GDP scenarios | regime-switching models | stress tests

ABSTRACT

We propose an approach for generating macroeconomic density forecasts that incorporate information on multiple scenarios defined by experts. We adopt a regime-switching framework in which sets of scenarios ("views") are used as Bayesian priors on economic regimes. Predictive densities coming from different views are then combined by optimizing objective functions of density forecasting. We illustrate the approach with an empirical application to quarterly real-time forecasts of the US GDP growth rate, in which we exploit the Fed's macroeconomic scenarios used for bank stress tests. We show that the approach achieves good accuracy in terms of average predictive scores and good calibration of forecast distributions. Moreover, it can be used to evaluate the contribution of economists' scenarios to density forecast performance.

Jel Classification: C11, C13, C22, C53

1 | Introduction

In recent years, it has become increasingly important for economic agents and policymakers to take account of the uncertainty around macroeconomic outlooks. In particular, two distinct approaches to this issue have emerged. The first one consists in generating economic forecasts in the form of (continuous) probability distributions, or *density forecasts*. This is now common practice among forecasters (Elliott and Timmermann 2016), and many economic agents, such as financial institutions, routinely evaluate their potential losses as random draws from predictive density functions (e.g., for the computation of value at risk and expected shortfall, see Jorion 2006). The second approach is to define a small number of (discrete) macroeconomic scenarios (e.g., see Moody's 2018). This approach facilitates communication to the public regarding economic uncertainty and finds important applications in financial supervision and risk management, most notably in the design of bank stress tests, which are now integral part of the financial regulatory framework in advanced economies (e.g., Federal Reserve Board 2018).

This paper develops a novel forecasting approach that combines these two perspectives. Using a regime-switching model, we construct macroeconomic density forecasts that explicitly incorporate information on discrete scenarios defined by economists. We consider alternative sets of scenarios, which we label as *views*, and use them to define Bayesian priors on economic regimes in the model. Next, as different views result in different density forecasts, we combine these forecasts in a way that optimizes standard evaluation criteria of density forecasting. The approach is illustrated through an empirical application to US GDP growth forecasts, in which we exploit the information contained in the macroeconomic scenarios defined by the Federal Reserve for its bank stress tests. We show that the approach achieves good forecast accuracy and good calibration of forecast distributions. Moreover, we find that its forecast performance is further improved when we combine it with the beta transformation of linear opinion pools proposed by Gneiting and Ranjan (2010, 2013). This nonlinear transformation allows to further recalibrate the optimally weighted linear combinations of density forecasts. Thus, we can use it to provide a nonlinear extension of our forecast aggregation approach.

On the one hand, the proposed approach is able to produce highly flexible predictive distributions. Such flexibility results both from the regime-switching framework and from the forecast combination procedure. Density forecasts from regime-switching models are mixture distributions, by construction: they are weighted averages of regime-specific densities, where the weights consist in the probabilities of the economy ending up in the various regimes. Likewise, density forecast combinations also produce mixture distributions. Thus, the approach can easily accommodate a wide range of departures from normality, as mixture distributions are in general nonnormal even if their individual components are. This is an important property when it comes to forecasting macroeconomic variables, whose empirical distributions often deviate from Gaussianity. In particular, a number of studies, including Fagiolo, Napoletano, and Roventini (2008), Cúrdia, Del Negro, and Greenwald (2014), and Ascari, Fagiolo, and Roventini (2015), have documented a nonnormal distribution of the GDP growth rate in the United States and other advanced economies, mainly as a result of large downturns. Acemoglu, Ozdaglar, and Tahbaz-Salehi (2017) find similar results and propose a theoretical model in which systematic departures of output from the normal distribution are explained by the interaction of idiosyncratic microeconomic shocks and sectoral heterogeneity. Compared to econometric approaches that impose nonnormal distributions of errors (e.g., Hansen 1994), regime-switching models allow for a more transparent economic explanation of nonnormality, based on the transition between different regimes.

On the other hand, regime-switching models appear as a natural framework for dealing with discrete scenarios such as those used in bank stress tests. First of all, these scenarios are typically constructed in a way that directly relates to economic regimes. For instance, the Fed defines its scenarios as "sets of economic and financial conditions", with baseline scenarios reflecting the most likely conditions and adverse/severely adverse scenarios reflecting conditions that prevail in recessions (Federal Reserve 2013, Appendix A of Part 252).[1] Also, many contributions in the literature on stress testing call for an important role of regime-switching models in the design of macro scenarios, arguing that realistic scenarios should reflect the nonlinearities that derive from the state dependency of macrofinancial conditions (Adrian, Morsink, and Schumacher 2020; Han and Leika 2019; Borio, Drehmann, and Tsatsaronis 2014; Gross, Henry, and Rancoita 2022; Bidder and McKenna 2015). As highlighted by Hamilton (2016), Markov-switching models provide a sufficiently parsimonious and yet robust characterization of the transition into and out of recession regimes, at least for the US economy.

This paper contributes to several strands of the forecasting literature. First, it relates to other papers that incorporate external—possibly judgmental—forecasts into econometric model-based forecasts. Krüger, Clark, and Ravazzolo (2017) combine forecasts from Bayesian VARs (BVARs) with forecasts from other sources using entropic tilting, a method for modifying a distribution in a way that satisfies specific moment conditions (Robertson, Tallman, and Whiteman 2005). Faust and Wright (2009) use the Federal Reserve Board's Greenbook forecasts as data in autoregressive and factor-augmented autoregressive models, in order to produce forecasts of GDP growth and inflation. Schorfheide and Song (2015) and Wolters (2015) also use nowcasts from Fed's Greenbook as data to generate forecasts from a BVAR and DSGE models, respectively. The main difference between our approach and these approaches is that we exploit information on multiple economic scenarios from external sources, mapping them into different regimes, while these approaches only target point forecasts or forecast variance.

Second, the paper relates to other contributions on forecasting using regime-switching models. The most closely related approaches are Bayesian ones, such as those by Pesaran, Pettenuzzo, and Timmermann (2006), who use a break point model (a generalization of regime-switching models) with hyperparameter uncertainty, and by Bauwens, Carpantier, and Dufays (2017), who estimate a Markov-switching model with an unknown and potentially infinite number of regimes. Unlike these studies, our paper exploits experts' views to define priors on economic regimes. A broad literature has investigated the forecast performance of regime-switching models. While the available evidence on the accuracy of point forecasts is mixed (Elliott and Timmermann 2016), these models have proved very useful for density forecasting and forecasts of tail events. Chauvet and Potter (2013) show that Markov-switching models can achieve high accuracy in forecasting GDP, especially with respect to the timing and depth of recessions. Geweke and Amisano (2011) show the usefulness of Markov mixtures for density forecasts of the stock market. Bauwens, Carpantier, and Dufays (2017) use an infinite Markov-switching autoregressive moving average model to produce density forecasts of GDP. Alessandri and Mumtaz (2017) find that a threshold VAR in which regime shifts depend on financial conditions produces good density forecasts of US GDP during the Great Recession.

Third, the paper relates to the large literature on forecast combinations. In particular, it has been shown that gains in density forecast performance are often achieved by combining different predictive distributions. Contributions in the field of macroeconomic forecasting include Hall and Mitchell (2007), Geweke and Amisano (2011), and Ganics (2017), among others. A vast body of literature in other fields, such as management science, risk analysis, and meteorology, also deals with the combination of probabilistic forecasts from experts into "consensus" distributions (e.g., Genest and Zidek 1986; Clemen and Winkler 1999). For comprehensive reviews, see Aastveit et al. (2018) and Wang et al. (2023). Several ways to further enhance density forecasts of nonnormal variables have been proposed in the literature on forecast combinations, such as beta opinion pools (Gneiting and Ranjan 2010, 2013) and empirically transformed opinion pools (Garratt, Henckel, and Vahey 2023), which will be explored in the empirical part of the paper.

We combine forecasts from different experts' views using two main criteria of density forecast evaluation. The first one is the sum of log scores, which measures the ability to assign high probabilities to outcomes that are truly likely to be observed. The second is a uniformity test on the probability integral transforms (PITs) of the forecasts, which measures the degree of calibration of the forecast distribution (Diebold, Gunther, and Tay 1998).[2] Both measures have been used in the literature to compute "optimal" forecast combinations (Hall and Mitchell 2007; Geweke and Amisano 2011; Ganics 2017).

The remainder of the paper is organized as follows: Section 2 explains the methodology, Section 3 presents the empirical application, and Section 4 concludes.

2 | Methodology

2.1 | The Regime-Switching Model

Our reference model is a Markov-switching autoregressive (MSAR) model in which the intercept and the error variance depend on the unobserved state of the economy. Let y_t denote a macroeconomic variable of interest at time t. The MSAR can be written as follows:

$$y_t = \sum_{j=1}^{p} \alpha_j y_{t-j} + \beta_{S_t} + \varepsilon_t \tag{1}$$

$$\varepsilon_t \sim \mathcal{N}\left(0, \sigma^2_{S_t}\right)$$

where S_t denotes the unobserved state variable at time t, β_{S_t} is the intercept in regime S_t, α_j for $j = 1, \ldots, p$ is a state-independent autoregressive coefficient, p is the maximum lag, ε_t is the error term, $\sigma^2_{S_t}$ is the regime-dependent variance of the error, and $\mathcal{N}$ denotes the normal distribution.[3] In particular, S_t is a Markov chain characterized by a transition matrix $\boldsymbol{\xi}$, where the element ξ_{kj} in row k and column j represents the probability of transition from state k to state j:

$$\xi_{kj} = Pr(S_t = j | S_{t-1} = k) \tag{2}$$

with $k, j = 1, \ldots, K$, where K is the number of regimes in the economy. Accordingly, the MSAR captures the typical autocorrelation of macro variables in two ways: by means of the autoregressive coefficients in (1) and through the persistence in the state variable S_t, as expressed by the transition matrix. Finally, let $\boldsymbol{\vartheta}$ denote the vector of parameters of the MSAR model, that is, $\boldsymbol{\vartheta} = (\beta_1, \ldots, \beta_K, \sigma_1, \ldots, \sigma_K, \alpha_1, \ldots, \alpha_p, \boldsymbol{\xi})$.

2.2 | Incorporating Information on Discrete Scenarios

The MSAR model (1) is estimated using Bayesian methods (see the Supporting Information Appendix for details). The priors on the parameters follow conventional distributions (Frühwirth-Schnatter 2006), which are the following:

$$\alpha_j \sim \mathcal{N}\left(a_{j,0}, A_{j,0}\right) \tag{3}$$

$$\beta_k \sim \mathcal{N}\left(b_{0,k}, B_{0,k}\right) \tag{4}$$

$$\sigma^2_k \sim \mathcal{G}^{-1}\left(c_0, C_0\right) \tag{5}$$

for $j = 1, \ldots, p$, where $\mathcal{G}^{-1}$ denotes the inverse gamma distribution and $a_{j,0}, A_{j,0}, b_{0,k}, B_{0,k}, c_0, C_0$ are hyperparameters to be selected by the researcher. In addition, for the transition matrix $\boldsymbol{\xi}$, it is assumed that the rows are independent and each row follows a Dirichlet distribution, denoted by $\mathcal{D}$, that is,

$$\boldsymbol{\xi_k} \sim \mathcal{D}\left(e_{k1}, \ldots, e_{kK}\right) \tag{6}$$

where $e_{k1}, \ldots, e_{kK}$ are hyperparameters, for $k = 1, \ldots, K$.

We define a *view* as a set of assumptions concerning the number of regimes K and the model parameters, that is, a specific set of values for K and for the hyperparameters of the model's prior.

Views can incorporate information on discrete economic scenarios defined by experts. In particular, we consider priors in which each regime is "centered" on a corresponding scenario using the following rule. Consider an AR(p) model where the coefficients are given by the k-state-specific prior hyperparameters, that is,

$$y_t = \sum_{j=1}^{p} a_{j,0} y_{t-j} + b_{0,k} + \varepsilon_t \tag{7}$$

In this model, the unconditional expectation of y_t, which we denote by $E_k(y_t)$, is as follows:

$$E_k(y_t) = \frac{b_{0,k}}{1 - \sum_{j=1}^{p} a_{j,0}} \tag{8}$$

Then, given an assumption on the state-independent autoregressive coefficients a_j, each regime-specific hyperparameter $b_{0,k}$ is chosen in such a way that expectation (8) matches a specific value derived from a scenario provided by external sources. The hyperparameter $B_{0,k}$ determines the tightness of the prior around these expectations or, in other words, the strength of economists' views. In principle, all the other hyperparameters can also be set to values suggested by economists, if available. Otherwise, uninformative or diffuse priors can be used for the remaining parameters of the MSAR.

When several alternative views are considered, Bayesian averaging across views can be performed. Let $\boldsymbol{\vartheta}^0_{K,i}$ denote the generic i-th view assuming K states and let $\boldsymbol{\pi}^0$ be a vector containing discrete prior probabilities assigned to all views considered. The posterior probability of view $\boldsymbol{\vartheta}^0_{K,i}$, which we denote as $\pi_{K,i}$, will depend on the prior probability vector $\boldsymbol{\pi}^0$ and on the values of the marginal likelihood of the MSAR model associated with the different views. Note that the letter π is used throughout the text to denote discrete probability distributions.

Further details are provided in the Supporting Information Appendix.

2.3 | Optimizing Density Forecasts

Let us define the vector containing all observations up to time t as $\mathbf{y}_t$, i.e., $\mathbf{y}_t = (y_1, \ldots, y_t)$, and let us consider a forecast horizon of one period. Then, let us denote the density forecast produced by any given view at time t as $p\left(y_{t+1} | \mathbf{y}_t, \boldsymbol{\vartheta}^0_{K,i}\right)$ (see Supporting Information Appendix A.2 for details on how density forecasts are calculated). Fixing a maximum number of regimes $\overline{K}$, for any $K = 1, \ldots, \overline{K}$ we consider a number P_K of alternative views assuming K regimes in the economy. We consider two methods for pooling forecasts across different views:

- *Forecast combinations*. We can express a forecast combination across MSAR views, where the vector of combination weights is denoted by $\mathbf{w}$, as follows:

$$p\left(y_{t+1}|\mathbf{y}_t,\mathbf{w}\right)=\sum_{K=1}^{\overline{K}}\sum_{i=1}^{P_K}p\left(y_{t+1}|\mathbf{y}_t,\boldsymbol{\vartheta}_{K,i}^0\right)w_{K,i} \tag{9}$$

where $w_{K,i}\geq 0$ is the weight assigned to view $\boldsymbol{\vartheta}_{K,i}^0$, with $\sum_{K=1}^{\overline{K}}\sum_{i=1}^{P_K}w_{K,i}=1$.

- *Bayesian averaging* over different views, that is,

$$p\left(y_{t+1}|\mathbf{y}_t,\boldsymbol{\pi}^0\right)=\sum_{K=1}^{\overline{K}}\sum_{i=1}^{P_K}p\left(y_{t+1}|\mathbf{y}_t,\boldsymbol{\vartheta}_{K,i}^0\right)\pi_{K,i} \tag{10}$$

Forecast (10) is a weighted average in which the weight assigned to the view-specific forecast $p\left(y_{t+1}|\mathbf{y}_t,\boldsymbol{\vartheta}_{K,i}^0\right)$ is given by the posterior probability of view $\boldsymbol{\vartheta}_{K,i}^0$, that is, $\pi_{K,i}$, which depends on the prior probability vector $\boldsymbol{\pi}^0$, with $\sum_{K=1}^{\overline{K}}\sum_{i=1}^{P_K}\pi_{K,i}=1$ (see the Supporting Information Appendix for more details).

We find the "optimal" weights $\mathbf{w}$ and priors $\boldsymbol{\pi}^0$ by maximizing two alternative objective functions, based on statistics that are commonly used to evaluate density forecast performance:

1. *Log scores*. The log score is the log of the predictive density function evaluated at the actual realization of the target variable. Given a generic forecast horizon h, let y_{t+h}^o (where o stands for "observed") denote the realization of variable y at time $t+h$, which is not observed at time t, when the forecast for $t+h$ is produced. Also, let R be the length of the timespan over which forecasts are optimized. The first objective function, denoted by f_1, is given by the sum of log scores over the period of interest. Specifically, our first objective function at a given time τ is as follows:

$$f_{1,\tau}(\boldsymbol{\omega})=\sum_{t=\tau-h-R+1}^{\tau-h}\ln\left(p\left(y_{t+h}^o|\mathbf{y}_t,\boldsymbol{\omega}\right)\right) \tag{11}$$

where $\boldsymbol{\omega}$ is a place-holder for either $\mathbf{w}$ or $\boldsymbol{\pi}^0$, depending on whether we consider forecast combinations or Bayesian averaging.

2. *PITs*. The PIT is the cumulative predictive density function evaluated at the actual realization of the variable. If the density forecast used to compute the PIT corresponds to the true distribution of the variable, then, for $h=1$, the PIT values are the realizations of independently and identically distributed (i.i.d.) uniform (0,1) variables (Diebold, Gunther, and Tay 1998). Therefore, a uniformity test on the PITs can be seen as a test of correct specification of the density forecasts (see also Rossi and Sekhposyan 2014). Accordingly, the second objective function is given by the following:

$$f_{2,\tau}(\boldsymbol{\omega})=-ks\left(\left\{\Phi\left(y_{t+1}^o|\mathbf{y}_t,\boldsymbol{\omega}\right)\right\}_{t=\tau-R}^{\tau-1}\right) \tag{12}$$

where $\Phi(\cdot)$ denotes the cumulative predictive density function, $ks(\cdot)$ indicates the function returning the test statistics of the Kolmogorov–Smirnov (KS) test of uniformity, and as before, $\boldsymbol{\omega}$ is a place-holder for either $\mathbf{w}$ or $\boldsymbol{\pi}^0$. Maximizing $-ks(\cdot)$ is equivalent to maximizing the p-value of the KS test (whose null hypothesis is uniformity).

Both the optimization based on f_1 and the one based on f_2 are solved numerically. For each f_i, with $i=1,2$, the optimization algorithm delivers two vectors at time τ: the vector of optimal forecast weights $\mathbf{w}_{i,\tau}^*$ for the set of alternative views, that is,

$$\mathbf{w}_{i,\tau}^*=\arg\max_{\mathbf{w}} f_{i,\tau}(\mathbf{w}) \tag{13}$$

and the vector of optimal prior probabilities $\boldsymbol{\pi}_{i,\tau}^{0*}$:

$$\boldsymbol{\pi}_{i,\tau}^{0*}=\arg\max_{\boldsymbol{\pi}^0} f_{i,\tau}\left(\boldsymbol{\pi}^0\right) \tag{14}$$

The former represents the typical problem explored in the literature on density forecast combination, whereas the latter can be seen as an empirical method for eliciting priors in the context of Bayesian model averaging. The optimal prior $\boldsymbol{\pi}_{i,\tau}^{0*}$ represents the discrete prior probability distribution of views such that the resulting posterior $\boldsymbol{\pi}_{i,\tau}^*$, when used as a vector of forecast weights, maximizes the density forecast performance, based on the selected objective function. In practice, the main difference between (13) and (14) is that the first problem directly delivers weights for forecast combination, while in the second case, the actual forecast weights are the posterior probabilities and so will also depend on the marginal likelihoods of all views, that is, $p(\mathbf{y}|\boldsymbol{\vartheta}_{K,i}^0)\ \forall K,i$ (see the Supporting Information Appendix).

3 | Empirical Application: US GDP Growth

This section illustrates the approach through an empirical application to density forecasts of US GDP growth. We use real-time quarterly data from 1948Q1 to 2019Q4[4] and consider the year-on-year growth rate of real GDP (expressed in percentage points in what follows).[5] Over this period, the (unconditional) distribution of GDP growth is nonnormal (the Jarque-Bera test rejects normality at the 1% level of significance) and, in particular, fat-tailed (kurtosis is close to 4), in line with the findings of prior literature (e.g., Fagiolo, Napoletano, and Roventini 2008; Ascari, Fagiolo, and Roventini 2015).

We set the lag length p of the model to 5, in consideration of the quarterly frequency of the variable. As explained below in more detail, we first produce pseudo-out-of-sample forecasts using a recursive window scheme. Then, we calculate optimal weights $\mathbf{w}^*$ and priors $\boldsymbol{\pi}^{0*}$ over time, following the procedure from Section 2.3. Finally, we assess the performance of pooled forecasts on an evaluation sample, that is, using observations of the target variable that have not been used in the optimization procedure.

3.1 | Priors and Fed Scenarios

We consider a total of 13 alternative views on the US GDP growth regimes. Eight views impose strongly informative priors derived from the scenarios of the Fed stress tests

2015–2018.[6] The remaining five views are "vague" views, defined by imposing uninformative priors on MSAR parameters. We consider different assumptions on the number of regimes $K = 1,2,3,4,5$.[7]

Let us first consider the Fed-based views. For each of the four stress tests under consideration, two views are constructed, one with $K = 3$ and the other with $K = 5$. In the view with $K = 3$, one of the regimes (which may be called the "normal times" regime) is derived from the Fed baseline scenario, another ("recession regime") from the adverse scenario, and the last one ("severe recession regime") from the severely adverse scenario.[8] In particular, we "center" each regime on the corresponding Fed scenario using the rule described in Section 2.2. Specifically, we consider an AR(5) model where the coefficients are given by the k-state-specific hyperparameters of the prior $\vartheta^0_{K,i}$, so Equation (8) implies $E_k(y_t) = b^{(K,i)}_{0,k} / \left(1 - \sum_{j=1}^{5} a^{(K,i)}_{j,0}\right)$. Then, after setting a value for the state-independent $a^{(K,i)}_{j,0}$, with $j = 1, \dots, 5$, each regime-specific $b^{(K,i)}_{0,k}$ is chosen in such a way that this expectation matches a specific value derived from the relevant scenario of the Fed stress test. For the normal times regime, this value is the average growth rate in the last four quarters of the baseline scenario, which is assumed to be close to the convergence value of the year-on-year growth rate in the absence of shocks.[9] For both the recession and the severe recession regimes, the value to be matched is the average growth rate in the first four quarters of the corresponding scenario, as the first quarters are those when the negative shocks are assumed to occur and the growth rates are lowest.

An example may help. Let us consider the view with $K = 3$ derived from the 2018 Fed stress test. The average growth rate of GDP in the last four quarters of the baseline scenario is 2.1%, while the average growth rates in the first four quarters of the adverse and severely adverse scenarios are −2.125% and −6.275%, respectively. As customary in the related Bayesian literature (e.g., Alessandri and Mumtaz 2017; Frühwirth-Schnatter 2006), we set the prior on the autoregressive component of the model using a parsimonious specification. In particular, we set the prior mean of the autoregressive coefficients to the (approximate) OLS estimate of an AR(1), that is, to 0.9 for the first lag and 0 for the higher-order lags (recall, however, that we allow up to 5 lags in the posterior estimates of the model). Then, $\sum_{j=1}^{5} a^{(K,i)}_{j,0} = 0.9$. Accordingly, the prior means for the regime-specific intercepts are set to $b_{0,1} = 2.1/(1-0.9) = 0.21$ for the normal times regime, $b_{0,2} = -2.125/(1-0.9) = -0.2125$ for the recession regime, and $b_{0,3} = -6.275/(1-0.9) = -0.6275$ for the severe recession regime.

The four stress test-based views with $K = 5$ expand the views with $K = 3$ by adding two regimes: a regime which we may call "recovery from recession", designed to match the last four quarters of the adverse scenario, and a regime of "recovery from severe recession", which matches the last four quarters of the severely adverse scenario. This is done in consideration of the fact that growth rates in the last four quarters of Fed's adverse and severely scenarios are assumed to be higher than the baseline rates, implying a rebound of the economy after a negative shock. Of course, such regimes may be more generally interpreted as "favorable regimes" characterized by positive shocks and not necessarily as recoveries from recessions.

In the five vague views, all priors on the intercepts are centered on 0 and have a variance of 1 percentage point, while the priors on the autoregressive coefficients are centered on 0.5 for the first lag, on 0 for the higher-order lags, and have a variance of 1. Taken jointly, these assumptions imply a large prior variance on the regime-specific means of the GDP growth rate. Conversely, in the Fed-based views, the priors for both β and α are strongly informative, so as to ensure that the regime-specific means are tightly centered on the stress test values. In particular, both priors are assumed to have minimal variance, equal to 10^{-5}. For the autoregressive coefficients α, the prior mean is assumed to be 0.9 for the first lag and 0 for higher-order lags, as mentioned in the previous example.

No strong assumption is made regarding the regime-switching error variance σ^2_k. Instead, a diffuse hierarchical prior is assumed in all views, using a gamma hyper-prior for C_0, that is, $C_0 \sim \mathcal{G}(g_0, G_0)$, where g_0 and G_0 are hyperparameters.[10]

Finally, the hyperparameters for the k-th row of the transition matrix $\boldsymbol{\xi}$ are set in such a way that the (prior) expected probability of remaining in the same state k in the next period is $\mathrm{E}(\xi_{kk}) = 2/3$ regardless of the number of regimes K and for any k, while the probability of moving to a different, specific state j decreases with the number of regimes, specifically $\mathrm{E}(\xi_{kj}) = 1/[3(K-1)]$, for any $j \neq k$ (see Frühwirth-Schnatter 2006).[11]

The summary of the alternative views is provided in Table 1, where views 1–5 are the vague ones, while views 6–13 are those derived from the Fed stress tests 2015–2018. Table A1 in the Supporting Information Appendix reports the GDP scenarios of the Fed stress tests.

3.2 | Optimization Scheme

We generate a sequence of pseudo-out-of-sample density forecasts using a recursive-window estimation scheme.[12] Next, the forecasts are used to carry out the optimization of weights/priors, which is iterated over time. The procedure can be described as follows. Let us assume that we are at time T_w and the forecast horizon is h. For each view under consideration, the MSAR model is recursively estimated using observations between an initial time t_0 and time t, with $t = T_0, T_0 + 1, \dots, T_w - h$. Time T_0 is therefore the end period of the shortest estimation sample. Estimates at T_0 are used to make forecasts for period $T_0 + h$; estimates at $T_0 + 1$ are used to make forecasts for $T_0 + 1 + h$, and so on. Thus, at time T_w, a sequence of past forecasts is available for each view. At this point, the algorithm computes the optimal weights/priors based on the last R forecasts, that is, maximizes the relevant objective function between $T_w - R + 1$ and T_w. Once the optimal weights/priors are retrieved, they are used to combine the different view-specific forecasts for the future period $T_w + h$, which is out of the optimization sample. When the actual value of the variable of interest is observed, at time $T_w + h$, the performance of the combined forecast is measured. The index T_w

TABLE 1 | Alternative views for the regime-switching model of US GDP growth.

		Hyperparameters								
View no.	**View type**	K	b_0	B_0	a_0	A_0	e_{kk}	c_0	g_0	G_0
1	Vague	1	0	1	(0.5 0 0 0 0)	1	2	3	0.5	0.5
2	Vague	2	(0,0)	1	(0.5 0 0 0 0)	1	2	3	0.5	0.5
3	Vague	3	(0,0,0)	1	(0.5 0 0 0 0)	1	2	3	0.5	0.5
4	Vague	4	(0,0,0,0)	1	(0.5 0 0 0 0)	1	2	3	0.5	0.5
5	Vague	5	(0,0,0,0,0)	1	(0.5 0 0 0 0)	1	2	3	0.5	0.5
6	Fed stress test	3	(0.265,−0.0475,−0.4275)	10^{-5}	(0.9 0 0 0 0)	10^{-5}	2	3	0.5	0.5
7	Fed stress test	3	(0.2275,−0.1850,−0.5675)	10^{-5}	(0.9 0 0 0 0)	10^{-5}	2	3	0.5	0.5
8	Fed stress test	3	(0.205,−0.1950,−0.59)	10^{-5}	(0.9 0 0 0 0)	10^{-5}	2	3	0.5	0.5
9	Fed stress test	3	(0.21,−0.2125,−0.6275)	10^{-5}	(0.9 0 0 0 0)	10^{-5}	2	3	0.5	0.5
10	Fed stress test	5	(0.39, 0.1975, 0.265,−0.0475,−0.4275)	10^{-5}	(0.9 0 0 0 0)	10^{-5}	2	3	0.5	0.5
11	Fed stress test	5	(0.39, 0.3, 0.2275,−0.1850,−0.5675)	10^{-5}	(0.9 0 0 0 0)	10^{-5}	2	3	0.5	0.5
12	Fed stress test	5	(0.39, 0.3, 0.205,−0.1950,−0.59)	10^{-5}	(0.9 0 0 0 0)	10^{-5}	2	3	0.5	0.5
13	Fed stress test	5	(0.43, 0.32, 0.21,−0.2125,−0.6275)	10^{-5}	(0.9 0 0 0 0)	10^{-5}	2	3	0.5	0.5

Note: The table lists the 13 priors (*views*) used to estimate the Bayesian Markov-switching autoregressive (MSAR) model considered in the empirical application. K denotes the assumed number of regimes; b_0 , B_0 , a_0 , A_0 , e_{kk} , c_0 , g_0, and G_0 are the hyperparameters of the priors. Please refer to Section 2 for an explanation of the parameters. Views 1–5 represent diffuse priors, while views 6-13 are strongly informative priors derived from the Fed supervisory scenarios.

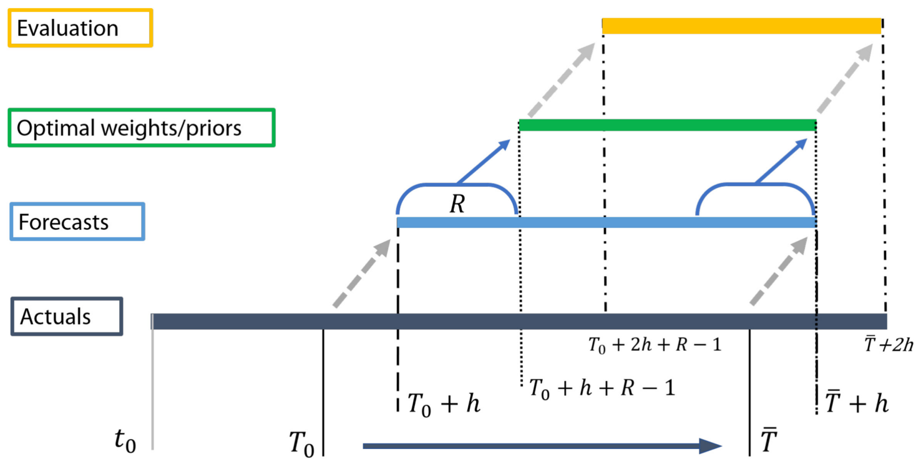

FIGURE 1 | Optimization scheme. *Note:* The figure summarizes the density forecast optimization scheme. First, the MSAR model is recursively estimated on actual GDP data (dark blue bar) using alternative views. The sample start date is denoted with t_0; the end date runs from T_0 to $\overline{T}$. For each sample window, the estimates generate density forecasts with horizon h (light blue bar). A rolling sequence of R forecasts is used to compute optimal forecast weights and prior probabilities (green bar) for the views. The optimal weights/priors obtained in each period are used to combine the view-specific forecasts for subsequent periods. The resulting composite forecasts (dark yellow bar) are evaluated by comparison with the actual data over the period from $T_0 + 2h + R - 1$ to $\overline{T} + 2h$.

runs from $T_0 + h + R - 1$ to $\overline{T} + h$, where $\overline{T}$ is the end of the largest estimation sample. $\overline{T} + 2h$ is the last available observation for the target variable. Therefore, the period from $T_0 + 2h + R - 1$ to $\overline{T} + 2h$ defines the evaluation sample, or test set. Figure 1 summarizes the procedure, which closely follows Ganics (2017).

More specifically, the application to the US GDP growth sets $t_0 = 1948Q1$, $T_0 = 1967Q4$, $R = 40$ quarters, $h = 1$ quarter, and $\overline{T} = 2019Q2$. Accordingly, the evaluation sample runs from 1978Q1 to 2019Q4.[13] We focus on the forecast horizon $h = 1$ because, as previously mentioned, the result of PIT uniformity of well-behaved forecasts, used for the optimization, only holds for $h = 1$ (see Rossi and Sekhposyan 2014).

3.3 | Results

3.3.1 | Main Results

Table 2 shows the performance of our composite regime-switching forecasts using optimal forecast weights and optimal priors and compares it with benchmark approaches. We label our forecasts as scenario-augmented MSAR (SA-MSAR) forecasts. As mentioned in Section 2.3, weights $\mathbf{w}_1^*$ and priors $\boldsymbol{\pi}_1^*$ result from the optimization based on the sum of log scores, while $\mathbf{w}_2^*$ and $\boldsymbol{\pi}_2^*$ are obtained by maximizing the p-value of the KS test of uniformity for the PITs. The first benchmark considered is a simple AR model (corresponding to view 1 in Table 1). Next, we consider three models that allow for nonnormal and heteroskedastic errors: an AR with Student-t errors, an AR with ARCH errors, and an AR with GARCH errors. As with the MSAR models, the lag length for the AR component is set to 5 for all models, while the ARCH and GARCH components have a lag length of 1. The remaining two benchmarks consist in uniform combination schemes over the alternative views, assigning equal forecast weights/prior probabilities to different values of K and, for any given K, equal weights/probabilities to the alternative views defined using K regimes.[14]

TABLE 2 | Density forecast performance.

Forecasting method	APD	KS
AR	0.28	0.00
AR-t	0.31	0.00
AR-ARCH	0.30	0.00
AR-GARCH	0.21	0.00
SA-MSAR - Equal forecast weights	0.32	0.00
SA-MSAR - Equal prior probabilities	0.36	0.00
SA-MSAR - Optimal weights $\mathbf{w}_1^*$	**0.37**	**0.00**
SA-MSAR - Optimal priors $\boldsymbol{\pi}_1^*$	**0.37**	**0.00**
SA-MSAR - Optimal weights $\mathbf{w}_2^*$	**0.33**	**0.07**
SA-MSAR - Optimal priors $\boldsymbol{\pi}_2^*$	**0.34**	**0.06**

Note: The table reports the density forecast performance of our scenario-augmented Markov-switching autoregressive (SA-MSAR) model for US GDP using optimal pools of views and compares it with several benchmarks. The optimal pools include log-score-based forecast combinations (optimal weights $\mathbf{w}_1^*$), log-score-based Bayesian averaging (optimal prior probabilities $\boldsymbol{\pi}_1^*$), PIT-based forecast combinations (optimal weights $\mathbf{w}_2^*$), and PIT-based Bayesian averaging (optimal prior probabilities $\boldsymbol{\pi}_2^*$). APD denotes the average predictive density; KS denotes the p-value of the Kolmogorov–Smirnov test of uniformity of the PITs. All statistics are computed over the period 1978Q1-2019Q4. Values in bold emphasis highlight the methodology proposed in this paper, as opposed to several benchmark methods (not in bold).

The table shows the average predictive density (APD), that is, the average of the exponential of log scores, and the p-value of the KS test.

The results indicate that our optimized SA-MSAR forecasts achieve good forecast accuracy and good calibration of forecast distributions. When optimized by log scores, the SA-MSAR forecasts outperform all benchmarks in terms of APDs. When optimized by PITs, they lead to nonrejection of the null hypothesis of uniformity at the 5% level in the KS test, indicating a reliable specification of the conditional predictive distribution of GDP growth. Thus, optimized SA-MSAR forecasts achieve well-behaved PITs.[15] By contrast, all benchmarks lead to rejection of the PIT uniformity hypothesis.

The approach can be used to evaluate the time-varying contribution of different views to the composite forecasts. Figures 2 and 3 display the evolution over time of the optimal forecast weights. The corresponding figures for the posterior probabilities resulting from the optimal priors are reported in the Supporting Information Appendix (Figures A1 and A2). In each figure, the area chart in the left panel shows the time-varying weights for all views from 1978Q1 to 2019Q4. The right panel plots the cumulative weight assigned to the views derived from the Fed supervisory scenarios. Figure 2 shows the results of the optimization based on log scores, while Figure 3 shows the results of the optimization based on the PITs. As can be seen from Figure 2, the vague views tend to dominate in the case of log-score optimization. In terms of optimal weights $\mathbf{w}_1^*$, the cumulative weight of the Fed-based views lies in the range 7%–42% until 1990 and is zero afterwards. Similarly, the Fed-based views provide nonzero contributions to optimized posteriors only between 1982 and 1987 (see Figure A1 in the Supporting Information Appendix).

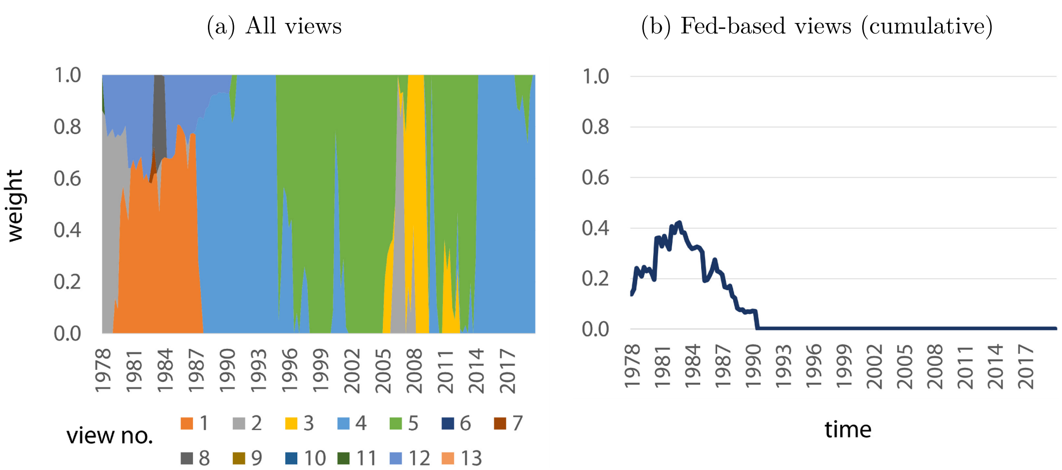

FIGURE 2 | Optimal log-score-based forecast combination weights ($\mathbf{w}_1^*$) over time. *Note:* The area chart in the left panel shows the time-varying forecast combination weights for all the 13 views used to estimate the Markov-switching AR model. The chart goes from 1978Q1 to 2019Q4. The weights ($\mathbf{w}_1^*$) are obtained using the log-score-based optimization procedure. The right panel plots the cumulative weight assigned to the views derived from the Fed's supervisory scenarios (views 6-13). See Table 1 for the list of views.

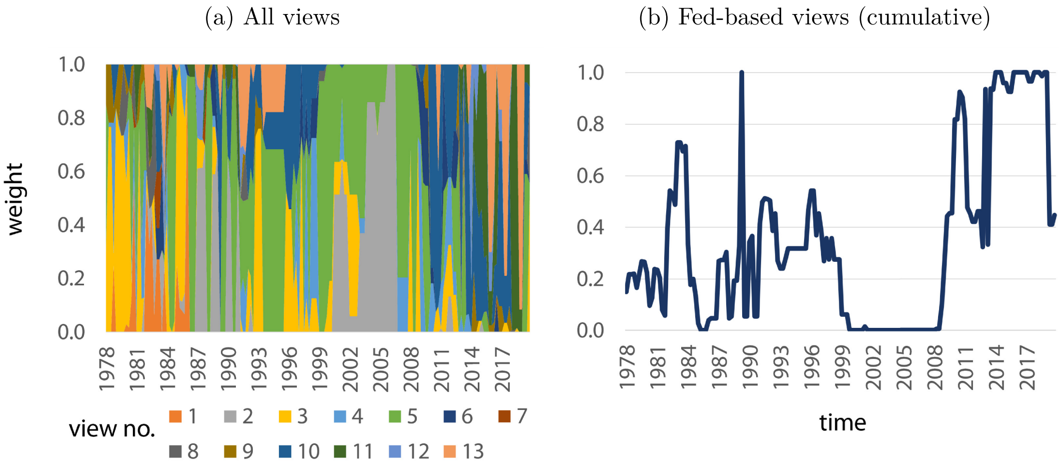

FIGURE 3 | Optimal PIT-based forecast combination weights ($\mathbf{w}_2^*$) over time. *Note:* The area chart in the left panel shows the time-varying forecast combination weights for all the 13 views used to estimate the Markov-switching AR model. The chart goes from 1978Q1 to 2019Q4. The weights ($\mathbf{w}_2^*$) are obtained using the PIT-based optimization procedure. The right panel plots the cumulative weight assigned to the views derived from the Fed's supervisory scenarios (views 6-13). See Table 1 for the list of views.

Overall, these results indicate a minor role of Fed-based views in boosting density forecast accuracy.

However, when the PIT-based optimization is considered, the contribution of the Fed-based views is much greater. On average, they account for about 35% of the combined forecasts in the case of optimal weights (Figure 2) and 27% in the case of posterior probabilities (see Figure A2 in the Supporting Information Appendix). In particular, they dominate in the period following the Global Financial Crisis, in terms of both $\mathbf{w}_2^*$ and $\boldsymbol{\pi}_2^*$. They also provide a substantial contribution in the early 1980s and during the 1990s. It is important to remark that using a single view is not sufficient to achieve well-calibrated forecasts. None of these views, when considered individually, leads to nonrejection of the PIT uniformity hypothesis in the KS test. Instead, the combination of different views is what drives the good results in terms of calibration.

To illustrate the outcome of our procedure in more detail, Figure 4 shows density forecasts for a specific quarter, 2016Q4, taken as an example. The figure displays three probability density functions (PDFs): the density forecast from an AR model (red line), the forecast generated by a three-regime MSAR model using a prior derived from the Fed stress test scenarios (green line; view 9 in Table 1), and the optimized MSAR forecast (blue line), that is, the final outcome of our procedure, using combination weights $\mathbf{w}_2^*$. Three dashed vertical lines indicate the three discrete scenarios for 2016Q4 contained in Fed's 2016 stress test. The figure provides an example of what can be obtained from our approach, compared to what is obtained from the single use of the two continuous/discrete approaches to forecast uncertainty. The (normal) density forecast of the AR model is a basic example of the classical density forecast outcome. The predictive density from the three-regime model is an example of scenario-augmented forecasts: it has a highly nonstandard left-skewed profile, clearly reflecting a mixture of three different regime-specific normals, strongly influenced by the tight prior centered on external scenarios. The final (blue) forecast is an example of the "optimal" combination of different densities (including the red and green ones reported in the figure). It has an irregular and leptokurtic

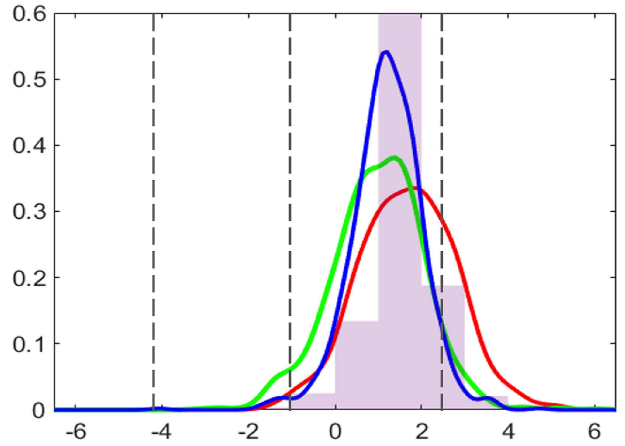

FIGURE 4 | Density forecasts for 2016Q4. *Note:* The figure shows three probability density functions (PDFs): the density forecast of GDP growth from an AR model (red line), the density forecast generated by a three-regime model using information from the Fed's stress test scenarios (green line, view 9 in Table 1), and the optimized scenario-augmented MSAR forecast (blue line), using combination weights $\mathbf{w}_2^*$. The three dashed vertical lines indicate the three scenarios for 2016Q4 contained in the 2016 Fed stress test. The figure also reports a histogram representing the distribution of the annual growth rate provided in 2016Q3 by the Survey of Professional Forecasters. The actual growth rate of GDP in 2016Q4 was 2%.

shape with two bumps in the tails, merging different forecasts as well as the prior information provided by external views. Finally, for comparison, the figure also reports a histogram representing the probability distribution of the annual growth rate provided in 2016Q3 (the forecast origin) by the Survey of Professional Forecasters.[16] Our "optimal" density forecasts are broadly in line with the SPF histogram but of course allow for a more detailed characterization of the forecast distribution.

Next, we examine the behavior of density forecasts across different parts of the GDP distribution. To this aim, Figure 5 plots the entire empirical distribution of the PITs, along with

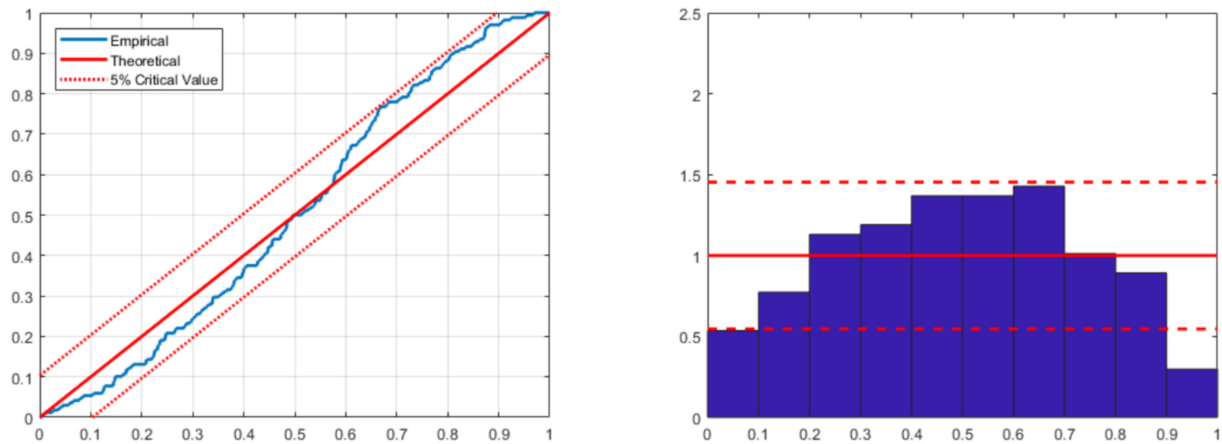

FIGURE 5 | Empirical distribution of the PITs. *Note:* The figure shows the empirical distribution of the probability integral transforms (PITs) associated with the optimized SA-MSAR density forecasts. The left panel shows the cumulative distribution function and the right panel shows the histogram of the PITs (normalized). In both panels, red solid lines indicate the uniform distribution and dashed lines indicate the 95% confidence interval of the uniform distribution.

the 95% confidence intervals of the uniform distribution, as calculated by Rossi and Sekhposyan (2019), which account for sample uncertainty.[17] The figure shows both the cumulative distribution function (CDF) and a histogram of the PITs (normalized), with deciles along the horizontal axis. For perfectly calibrated forecasts, the empirical CDF of the PITs would lie on the 45° line, and the histogram bars would all have a height of 1. We report the empirical distribution of the PITs associated with the optimized SA-MSAR forecasts using combination weights $\mathbf{w}_2^*$. As the figure shows, the empirical CDF always lies within the 95% confidence interval of the uniform distribution, in line with the result of the KS test in Table 2. In the histogram, all bars are within the confidence bands or on the bounds, except for the highest decile. This indicates that the density forecasts are well-behaved in most parts of the GDP growth distribution. Focusing on the tails, we note that the bar at the lowest decile lies approximately on the lower bound of the 95% confidence interval, suggesting that the forecasts capture left tail risk quite adequately. The bar at the highest decile lies below the lower bound, indicating some overdispersion of the forecasts in the upper tail (i.e., too few realizations of the target variable fall in the right tail of the forecast density).

3.3.2 | Extension: Applying Density Transformations

Lastly, we extend our approach by combining it with two types of density transformations that have been proposed in the literature to further improve density forecast performance: the beta-transformed linear pool (BLP; see Gneiting and Ranjan 2010, 2013) and the empirically transformed linear opinion pools (EtLOP, Garratt, Henckel, and Vahey 2023). The BLP applies a transformation of the conditional forecast densities using the beta distribution. This can improve the calibration and, in particular, the dispersion of density forecasts produced by simple linear combinations. The EtLOP transforms the density forecasts so that they match the empirical CDF of the target variable (GDP growth, in our case), in order to better accommodate non-Gaussianity.

TABLE 3 | Applying density transformations: BLP and EtLOP.

Forecasting method	APD	KS
EtLOP - AR	0.21	0.00
BLP - AR	0.39	0.18
EtLOP - SA-MSAR $\mathbf{w}_1^*$	0.26	0.00
EtLOP - SA-MSAR $\mathbf{w}_2^*$	0.26	0.00
BLP - SA-MSAR $\mathbf{w}_1^*$	**0.42**	**0.68**
BLP - SA-MSAR $\mathbf{w}_2^*$	**0.40**	**0.81**

Note: The table reports the average predictive density (APD) and the p-value of a Kolmogorov–Smirnov (KS) test of uniformity of the PITs associated with optimized scenario-augmented MSAR (SA-MSAR) forecasts, transformed using two alternative approaches: the beta-transformed linear pool (BLP; Gneiting and Ranjan 2010, 2013) and the empirically transformed linear opinion pools (EtLOP, Garratt, Henckel, and Vahey 2023). As a benchmark, the table also reports the results for BLP- and EtLOP-transformed AR forecasts. Values in bold emphasis highlight the extended version of our methodology which achieves the best results. The bold letter w in the other rows is motivated by the fact that the w variables are defined in bold in the text (see, e.g., Equation (9)).

We alternatively apply the beta transformation and the empirical transformation to the optimized linear combinations of density forecasts. In the case of BLP, we select the values of the two shape parameters of the beta distribution recursively, using the same optimization scheme described before. In the case of EtLOP, at any time t considered as the forecast origin, we transform the optimized density forecasts using the empirical distribution of GDP growth estimated nonparametrically (by kernel smoothing) on data up to time t.

Table 3 shows the results. We report results using optimal weights, but very similar conclusions are obtained using optimal priors. As a benchmark, we also report the results of BLP and EtLOP transformations applied to AR forecasts. As the table shows, the beta transformation further improves the density forecast performance of our scenario-augmented MSAR forecasts, in terms of both log scores and PITs. The ADP is now 0.4–0.42, and the p-value of the PIT uniformity test jumps to 0.68–0.81, depending on the specific weights used. Also, our

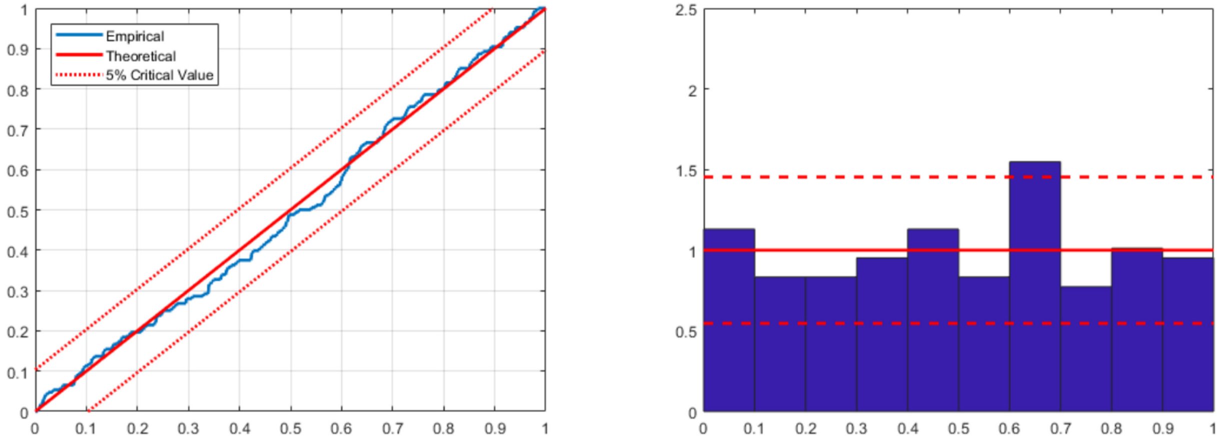

FIGURE 6 | Empirical distribution of the PITs: BLP-MSAR. *Note:* The figure shows the empirical distribution of the probability integral transforms (PITs) associated with the optimized SA-MSAR density forecasts, transformed using a beta distribution (Gneiting and Ranjan 2010, 2013). The left panel shows the cumulative distribution function and the right panel shows the histogram of the PITs (normalized). In both panels, red solid lines indicate the uniform distribution and dashed lines indicate the 95% confidence interval of the uniform distribution.

optimized SA-MSAR forecasts, once transformed by BLP, remain clearly superior to the benchmark BLP-transformed AR forecasts, in terms of both accuracy and calibration, although benchmark forecasts have also improved compared to their untransformed counterparts from Table 2. Conversely, the EtLOP methodology does not appear to provide any improvement in this specific empirical application.

Finally, Figure 6 shows the entire empirical distribution of the PITs in the case of beta-transformed optimal SA-MSAR forecasts (using combination weights $\mathbf{w}_2^*$). In this case, the highest decile also lies within the 95% confidence bands of the uniform distribution.

These results can be interpreted in light of the methodological differences between EtLOP and BLP. The main advantage of EtLOP is the great flexibility of the nonparametrically fitted empirical distribution used to transform the forecasts. However, the distribution can be fitted poorly, especially in the tails. Moreover, the approach constrains the forecast densities to align with the distribution of past observations. The BLP parametric approach is less flexible in adapting to empirical data but gives the forecaster more control over the shape of the forecasts, as they do not have to conform to the historical distribution of the target variable. In our empirical application, the fine-tuning of the beta transformation leads to concentrate more probability mass in the central part of the forecast distribution and less in the tails, compared to the untransformed forecasts. This adjustment further refines the calibration of the tails, as can be seen by comparing Figure 6 with Figure 5. Conversely, in this context, the empirically transformed distributions tend to underperform especially in the left tail (underestimating the frequency of realizations).

4 | Conclusions

We have developed an approach for generating density forecasts of macroeconomic variables using a variety of discrete economic scenarios provided by external sources. The approach is based on a Bayesian regime-switching model in which experts' views are translated into priors on economic regimes, and different views are pooled together to enhance density forecast performance.

We have presented an empirical application in which density forecasts of US GDP are formed using the supervisory scenarios defined by the Fed. We have shown that the approach is able to achieve both good forecast accuracy and correct calibration of predictive distributions, merging the flexibility of mixture predictive densities provided by regime-switching models with the benefits of forecast combination, which are well-established in the literature.

Importantly, this methodology allows to evaluate the usefulness of economists' views for density forecasting. To illustrate this possibility, the empirical application tracks the contribution of Fed's scenarios to the optimized forecasts over time.

This approach appears particularly valuable in all contexts in which tail risk has a clear economic interpretation and when economic projections have to comply with external, possibly judgmental views. Researchers and practitioners interested in this kind of analysis may fine-tune the approach by tailoring the range of views to be considered and by selecting different objective functions in the optimization procedure.

Acknowledgments

I thank Siem Jan Koopman (the editor) and two anonymous reviewers for very useful comments and suggestions. The usual disclaimer applies. Open access publishing facilitated by Universita degli Studi di Bologna, as part of the Wiley - CRUI-CARE agreement.

Data Availability Statement

The data that support the findings of this study are openly available in the Archival FRED database (https://alfred.stlouisfed.org/series?seid=GDPC1) and in the website of the Board of Governors of the Federal Reserve System (https://www.federalreserve.gov/supervisionreg/dfast-archive.htm).

Endnotes

[1] As explained in Federal Reserve (2013, Appendix A of Part 252), "the baseline scenario is designed to represent a consensus expectation" and is constructed using the forecasts of private sector forecasters (e.g., Blue Chip Consensus Forecasts and the Survey of Professional Forecasters), government agencies, and other public-sector organizations (e.g., the International Monetary Fund and the Organization for Economic Co-operation and Development). To develop the severely adverse scenario, the Fed relies on a recession approach, i.e., it specifies the "future paths of variables to reflect conditions that characterize post-war U.S. recessions". The severity of recessions is mainly determined by the level and increase of the unemployment rate, as well as the occurrence of a housing recession. In post-war US recessions that are classified as severe, GDP has dropped about 3.5 percent and unemployment has increased by a total of about 4 percentage points, on average (Federal Reserve 2013, Appendix A of Part 252, Table 1). The adverse scenario reflects "a set of economic and financial conditions that are more adverse than those associated with the baseline scenario but less severe than those associated with the severely adverse scenario".

[2] A well-calibrated forecast is one that does not make systematic errors: if p is the predicted probability assigned to a given random event, then that event should empirically occur with frequency p.

[3] We assume state-independent autoregressive coefficients (see, e.g., Hamilton 1989) to avoid the overfitting problems that easily arise in macro applications of regime-switching models. In particular, as explained in Hamilton (2016), "inference about parameters that only show up in regime i can only come from observations within that regime. With postwar quarterly data that would mean about 50 observations from which to estimate all the parameters operating during recessions. One or two parameters could be estimated fairly well, but overfitting is again a potential concern in models with many parameters. For this reason researchers may want to limit the focus to a few of the most important parameters that are likely to change, such as the intercept and the variance".

Also, as pointed out by Hamilton (2016), while there is no theoretical reason to assume that all regime-specific densities are normal, the same concerns of overfitting suggest using a normal distribution, which is defined by just two parameters. The assumption of normality is generally made in the literature on regime-switching models, with very few exceptions, such as Dueker (1997). Some empirical evidence may be used to corroborate this assumption. For instance, Acemoglu, Ozdaglar, and Tahbaz-Salehi (2017) find that, once large downturns are excluded from the data, the US GDP growth distribution appears to be well approximated by a normal.

[4] *Source:* The US Bureau of Economic Analysis, Real Gross Domestic Product (series code: GDPC1). We use the complete set of real-time data vintages provided in the Archival FRED (ALFRED) database compiled by the Federal Reserve Bank of St. Louis. The first vintage was released in December 1991. For all sample windows ending before 1991Q4, we use the data contained in the first available vintage. As vintages are generally released at a monthly frequency, for each quarter we take the last vintage released within that quarter.

[5] Our choice of using the year-on-year (YoY) rather than the quarter-on-quarter (QoQ) growth rate of GDP in the MSAR model follows the same reasoning as in Binder and Gross (2013): since the YoY rate is more persistent than the QoQ rate, changes in the mean and variance of the YoY rate are more likely to reflect the transition between different economic regimes compared to changes in the QoQ rate, which are more affected by noisy events specific to any given quarter.

[6] The scenarios are available at https://www.federalreserve.gov/supervisionreg/dfast-archive.htm.

[7] Our choice of the maximum number of regimes is supported by the results of Bauwens, Carpantier, and Dufays (2017). These authors develop a model that allows for an infinite number of regimes, using a nonparametric Dirichlet process. However, when estimating the model for US GDP growth (with different breaks for the mean and variance parameters), they find that the posterior probability of the number of regimes being at most 5 lies between 98% and 100% for the mean parameters and between 74% and 100% for the variance, depending on the prior used for estimation.

[8] Although the Fed stress test scenarios represent hypothetical paths and not forecasts, they are intended to be plausible even when severe. Therefore, they can legitimately be assigned predictive probabilities (see, e.g., Yuen 2013) and used to form density forecasts.

[9] The stress test scenarios are defined in terms of annualized quarter-on-quarter growth rates, so that averaging over the last four quarters approximates the year-on-year growth rate in the last quarter.

[10] To make the prior on σ_k^2 diffuse, the following values are selected for the hyperparameters: $c_0 = 3$, $g_0 = 0.5$ and $G_0 = 0.5$. These imply that σ_k^2 has a prior expected value of 0.5 percentage points of GDP and a high prior variance of 1.25 percentage points (see Section A.3 in the Supporting Information Appendix).

[11] Specifically, the hyperparameters for the k-th row of the transition matrix ξ are $e_{kk} = 2$ and $e_{kj} = 1/(K-1)$ if $k \neq j$, $\forall k, j$. Given the properties of the Dirichlet distribution, $\mathrm{E}(\xi_{kj}) = e_{kj} / \left(\sum_{l=1}^{K} e_{kl} \right)$.

[12] In this context, the choice of using expanding windows, as opposed to rolling windows, increases the probability that the variable "visits" most or all the regimes within the sample, thus greatly facilitating the estimation of the regime-switching model.

[13] We estimate the MSAR model using the MATLAB package `bayesf Version 2.0` by Frühwirth-Schnatter (2008). For each MSAR estimate, the Markov Chain Monte Carlo (MCMC) algorithm uses 1000 iterations as burn-in and 1000 iterations to store the results. Starting from the sample of forecasts produced by the MCMC algorithm, a complete probability density function is fitted using a normal kernel smoothing function.

[14] For instance, in the case of equal prior probabilities, it is assumed that $\pi_K^0 = 1/\overline{K}$ for each K and that $\pi(\boldsymbol{\vartheta}_{K,i}^0 | K) = 1/P_K$ for each view $\boldsymbol{\vartheta}_{K,i}^0$.

[15] Since correct calibration implies that the PITs are realizations of i.i.d U(0,1) variables, we also test for the serial independence of the PITs. Specifically, following Rossi and Sekhposyan (2014), we perform a Ljung–Box test of independence for both the first and the second moments of the PITs. Both tests do not reject the null hypothesis of serial independence, implying that forecasts are well-calibrated also in this respect.

[16] Note that the SPF does not provide a continuous density function of the quarterly GDP growth rate but only probabilities associated with discrete intervals of annual GDP growth (e.g., the probability that annual GDP growth in 2016 will be between 1% and 2%).

[17] For this analysis, we use Matlab code shared by Rossi and Sekhposyan (2019). As mentioned by Adrian, Boyarchenko, and Giannone (2019), confidence bands should be taken as approximations when forecasts are computed using an expanding estimation windows.

References

Aastveit, K. A., J. Mitchell, F. Ravazzolo, and H. van Dijk. 2018. "The Evolution of Forecast Density Combinations in Economics." (*Discussion Papers 18-069/III*): Tinbergen Institute. https://ideas.repec.org/p/tin/wpaper/20180069.html.

Acemoglu, D., A. Ozdaglar, and A. Tahbaz-Salehi. 2017. "Microeconomic Origins of Macroeconomic Tail Risks." *American Economic Review* 107, no. 1: 54–108. https://www.aeaweb.org/articles?id=10.1257/aer.20151086.

Adrian, M. T., M. J. Morsink, and M. L. B. Schumacher. 2020. "Stress Testing at the IMF." (*Departmental Papers 20/04*): International Monetary Fund. https://www.elibrary.imf.org/view/journals/087/2020/016/article-A001-en.xml.

Adrian, T., N. Boyarchenko, and D. Giannone. 2019. "Vulnerable Growth." *American Economic Review* 109, no. 4: 1263–1289. https://www.aeaweb.org/articles?id=10.1257/aer.20161923.

Alessandri, P., and H. Mumtaz. 2017. "Financial Conditions and Density Forecasts for US Output and Inflation." *Review of Economic Dynamics* 24: 66–78. https://www.sciencedirect.com/science/article/abs/pii/S1094202517300042.

Ascari, G., G. Fagiolo, and A. Roventini. 2015. "Fat-Tail Distributions and Business-Cycle Models." *Macroeconomic Dynamics* 19, no. 2: 465–476. https://www.cambridge.org/core/journals/macroeconomic-dynamics/article/abs/fattail-distributions-and-businesscycle-models/28294508BEB16B3CAB235D878F8EC62C?utm_campaign=shareaholic&utm_medium=copy_link&utm_source=bookmark.

Bauwens, L., J. F. Carpantier, and A. Dufays. 2017. "Autoregressive Moving Average Infinite Hidden Markov-Switching Models." *Journal of Business and Economic Statistics* 35, no. 2: 162–182. https://www.tandfonline.com/doi/full/10.1080/07350015.2015.1123636.

Bidder, R., and A. McKenna. 2015. "Robust Stress Testing." (*Working Paper Series 2015-13*): Federal Reserve Bank of San Francisco. https://ideas.repec.org/p/fip/fedfwp/2015-13.html.

Binder, M., and M. Gross. 2013. "Regime-Switching Global Vector Autoregressive Models." (*Working Paper Series 1569*): European Central Bank. https://ideas.repec.org/p/ecb/ecbwps/20131569.html.

Borio, C., M. Drehmann, and K. Tsatsaronis. 2014. "Stress-Testing Macro Stress Testing: Does It Live Up to Expectations?." *Journal of Financial Stability* 12: 3–15. https://www.sciencedirect.com/science/article/pii/S1572308913000454.

Chauvet, M., and S. Potter. 2013. "Forecasting Output." edited by G. Elliott, and A. Timmermann, *Handbook of Economic Forecasting*, Handbook of Economic Forecasting, Vol. 2, 141–194. Elsevier. https://www.sciencedirect.com/science/article/pii/B9780444536839000037.

Clemen, R. T., and R. L. Winkler. 1999. "Combining Probability Distributions From Experts In Risk Analysis." *Risk Analysis* 19, no. 2: 187–203. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1539-6924.1999.tb00399.x.

Cúrdia, V., M. Del Negro, and D. L. Greenwald. 2014. "Rare Shocks, Great Recessions." *Journal of Applied Econometrics* 29, no. 7: 1031–1052. https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2395.

Diebold, F. X., T. A. Gunther, and A. S. Tay. 1998. "Evaluating Density Forecasts with Applications to Financial Risk Management." *International Economic Review* 39, no. 4: 863–883. https://www.jstor.org/stable/2527342.

Dueker, M. J. 1997. "Strengthening the Case for the Yield Curve as a Predictor of U.S. Recessions." *Federal Reserve Bank of St. Louis Review* 79, no. 2: 41–51. https://fraser.stlouisfed.org/title/review-federal-reserve-bank-st-louis-820/march-april-1997-620685?page=37.

Elliott, G., and A. Timmermann. 2016. *Economic Forecasting*. Princeton University Press.

Fagiolo, G., M. Napoletano, and A. Roventini. 2008. "Are Output Growth-Rate Distributions Fat-Tailed? Some Evidence From OECD Countries." *Journal of Applied Econometrics* 23, no. 5: 639–669. https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.1003.

Faust, J., and J. H. Wright. 2009. "Comparing Greenbook and Reduced Form Forecasts Using a Large Realtime Dataset." *Journal of Business & Economic Statistics* 27, no. 4: 468–479. https://doi.org/10.1198/jbes.2009.07214.

Federal Reserve. 2013. "Policy Statement on the Scenario Design Framework for Stress Testing." *Federal Register* 78: 71435–71448. https://www.federalregister.gov/documents/2013/11/29/2013-27009/policy-statement-on-the-scenario-design-framework-for-stress-testing.

Federal Reserve Board. 2018. "2018 Supervisory Scenarios for Annual Stress Tests Required Under the Dodd-Frank Act Stress Testing Rules and the Capital Plan Rule." (*report*): Federal Reserve Board. https://www.federalreserve.gov/publications/files/bcreg20180201a1.pdf.

Frühwirth-Schnatter, S. 2006. *Finite Mixture and Markov Switching Models*, Springer Series in Statistics. New York: Springer. https://link.springer.com/book/10.1007/978-0-387-35768-3.

Frühwirth-Schnatter, S. 2008. *Finite Mixture and Markov Switching Models. Implementation in Matlab Using the Package Bayesf Version 2.0*, Springer Series in Statistics. New York: Springer. http://statmath.wu.ac.at/~fruehwirth/monographie/.

Ganics, G. 2017. "Optimal Density Forecast Combinations." (*Working Papers N. 1751*): Banco de España. https://www.bde.es/wbe/en/publicaciones/analisis-economico-investigacion/documentos-trabajo/optimal-density-forecast-combinations.html.

Garratt, A., T. Henckel, and S. P. Vahey. 2023. "Empirically-Transformed Linear Opinion Pools." *International Journal of Forecasting* 39, no. 2: 736–753. https://www.sciencedirect.com/science/article/pii/S0169207022000322.

Genest, C., and J. V. Zidek. 1986. "Combining Probability Distributions: A Critique and An Annotated Bibliography." *Statistical Science* 1, no. 1: 114–135. http://www.jstor.org/stable/2245510.

Geweke, J., and G. Amisano. 2011. "Optimal Prediction Pools." *Journal of Econometrics* 164, no. 1: 130–141. https://www.sciencedirect.com/science/article/pii/S0304407611000455.

Gneiting, T., and R. Ranjan. 2010. "Combining Probability Forecasts." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 72, no. 1: 71–91. https://doi.org/10.1111/j.1467-9868.2009.00726.x.

Gneiting, T., and R. Ranjan. 2013. "Combining Predictive Distributions." *Electronic Journal of Statistics* 7: 1747–1782. https://doi.org/10.1214/13-EJS823.

Gross, M., J. Henry, and E. Rancoita. 2022. "Macrofinancial Stress Test Scenario Design—For Banks and Beyond." In *Handbook of Financial Stress Testing*, edited by J. D. Farmer, A. M. Kleinnijenhuis, T. Schuermann, and T. Wetzer, 77–97. Cambridge University Press. https://www.cambridge.org/core/books/handbook-of-financial-stress-testing/macrofinancial-stress-test-scenario-designfor-banks-and-beyond/E7FAC3273389E926D68E36344109C438?utm_campaign=shareaholic&utm_medium=copy_link&utm_source=bookmark.

Hall, S. G., and J. Mitchell. 2007. "Combining Density Forecasts." *International Journal of Forecasting* 23, no. 1: 1–13. https://www.sciencedirect.com/science/article/pii/S0169207006000768.

Hamilton, J. D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica* 57, no. 2: 357–384. http://www.jstor.org/stable/1912559.

Hamilton, J. D. 2016. "Macroeconomic Regimes and Regime Shifts." In *Handbook of Macroeconomics*, edited by J. B. Taylor, and H. Uhlig, Handbook of Macroeconomics, Vol. 2, 163–201. Elsevier. https://ideas.repec.org/h/eee/macchp/v2-163.html.

Han, F., and M. Leika. 2019. "Integrating Solvency and Liquidity Stress Tests: The Use of Markov Regime-Switching Models." (*IMF Working Papers 2019/250*): International Monetary Fund. https://ideas.repec.org/p/imf/imfwpa/2019-250.html.

Hansen, B. E. 1994. "Autoregressive Conditional Density Estimation." *International Economic Review* 35, no. 3: 705–730. http://www.jstor.org/stable/2527081.

Jorion, P. 2006. *Value at Risk: The New Benchmark for Managing Financial Risk*. 3rd Edition. McGraw Hill, New York.

Krüger, F., T. E. Clark, and F. Ravazzolo. 2017. "Using Entropic Tilting to Combine BVAR Forecasts With External Nowcasts." *Journal of Business & Economic Statistics* 35, no. 3: 470–485. https://doi.org/10.1080/07350015.2015.1087856.

Moody's. 2018. "U.S. Macroeconomic Outlook Alternative Scenarios." (*Tech. rep.*): Moody's Analytics. https://www.economy.com.

Pesaran, M. H., D. Pettenuzzo, and A. Timmermann. 2006. "Forecasting Time Series Subject to Multiple Structural Breaks." *Review of Economic Studies* 73, no. 4: 1057–1084. https://doi.org/10.1111/j.1467-937X.2006.00408.x.

Robertson, J. C., E. W. Tallman, and C. H. Whiteman. 2005. "Forecasting Using Relative Entropy." *Journal of Money, Credit and Banking* 37, no. 3: 383–401. http://www.jstor.org/stable/3839160.

Rossi, B., and T. Sekhposyan. 2014. "Evaluating Predictive Densities of US Output Growth and Inflation in a Large Macroeconomic Data Set." *International Journal of Forecasting* 30, no. 3: 662–682. https://www.sciencedirect.com/science/article/pii/S0169207013000460.

Rossi, B., and T. Sekhposyan. 2019. "Alternative Tests for Correct Specification of Conditional Predictive Densities." *Journal of Econometrics* 208, no. 2: 638–657. https://www.sciencedirect.com/science/article/pii/S0304407618302197.

Schorfheide, F., and D. Song. 2015. "Real-Time Forecasting With a Mixed-Frequency VAR." *Journal of Business & Economic Statistics* 33, no. 3: 366–380. https://doi.org/10.1080/07350015.2014.954707.

Wang, X., R. J. Hyndman, F. Li, and Y. Kang. 2023. "Forecast Combinations: An Over 50-Year Review." *International Journal of Forecasting* 39, no. 4: 1518–1547. https://www.sciencedirect.com/science/article/pii/S0169207022001480.

Wolters, M. H. 2015. "Evaluating Point and Density Forecasts of DSGE Models." *Journal of Applied Econometrics* 30, no. 1: 74–96. https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2363.

Yuen, K. 2013. "Determining the Severity of Macroeconomic Stress Scenarios." (*Supervisory Staff Reports*): Board of Governors of the Federal Reserve System, December. https://www.federalreserve.gov/bankinforeg/determining-the-severity-of-macroeconomic-stress-scenarios.htm.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.