Empirically-transformed Linear Opinion Pools*

Anthony Garratt
(University of Warwick)

Timo Henckel
(ANU and CAMA)

Shaun P. Vahey
(University of Warwick and CAMA)

August 3, 2020

Abstract

Many studies have found that combining density forecasts improves predictive accuracy for macroeconomic variables. A prevalent approach known as the Linear Opinion Pool (LOP) averages the probabilistic assessments produced by individual experts to produce a combination density. In practice, the LOP combination density forecast often fails to match the distribution of the target variable. In this paper, we propose a computationally convenient transformation for a single target macroeconomic variable with a non-Gaussian distribution. Our methodology involves a Smirnov transform to reshape the LOP combination forecasts using a nonparametric kernel-smoothed empirical cumulative distribution function. We illustrate our methodology with an application examining quarterly real-time forecasts for U.S. inflation over an evaluation sample from 1990:1 to 2017:2. Our proposed methodology improves performance relative to a LOP combination by 8-9% in terms of both the root mean squared forecast error and the continuous ranked probability score.

JEL codes: C32; C53; E37
Keywords: Density forecast combination; Smirnov transform; inflation.

*We thank Todd Clark, Domenico Giannone, Dean Croushore, Kevin Lee, Craig Thamotheram, Liz Wakerly, Yunyi Zhang, Anastasia Allayioti and participants at the National Bank of Belgium Real-time Economics Conference 2019 for helpful comments and suggestions.

1 Introduction

Monetary policymakers often warn about the risks to inflation and other macroeconomic variables. The Federal Open Market Committee, like most central banks, discusses these risks explicitly. Some central banks, including the Bank of England and Norges Bank, publish (at times) asymmetric predictive densities for key variables. Among others, Smith and Vahey (2016) and Adrian et al. (2019) examine methods to quantify the risks to macroeconomic variables, drawing explicit attention to the non-Gaussianity of some U.S. macroeconomic variables.

As noted by Rossi (2019), a large and expanding literature examines the performance of density forecast combinations with variants of the Linear Opinion Pool (LOP) for macroeconomic variables. The LOP approach averages (pools) the density forecasts produced by individuals experts, as if they are probabilistic opinions. The FRB Philadelphia uses an equal-weighted LOP approach to aggregate the individual density forecasts from the respondents in the Survey of Professional Forecasters (SPF). The SPF LOP combination (conditional) density forecasts are, at times, non-Gaussian for inflation and other macroeconomic variables.

The LOP aggregation approach ensures that typically the shape of the combination density forecast is more flexible than the individual forecast densities being combined. However, that flexibility does not necessarily give a good match with the non-Gaussian features of the sample data. For example, even if the individual experts have correctly calibrated forecast densities, the equal-weighted LOP combination will not be correctly calibrated; see, Ranjan and Gneiting (2010).

In this paper, we propose a remedy to improve the matching of the LOP to the distribution of the target variable. Our methodology involves applying a

(modified) Smirnov transform to reshape the LOP combination forecasts using an (inverse) Empirical Cumulative Distribution Function (ECDF) fitted non-parametrically to the target variable. Our non-parametric approach allows us to match non-Gaussian features in the sample data for the target variable.

We illustrate our proposed methodology with an application involving forecasting U.S. inflation. We utilise an expert space similar to Garratt et al. (2011), where each expert uses a "misspecified" linear time series model with Gaussian errors subject to "uncertain instabilities" to produce "real-time" h-step ahead density forecasts. Jore et al. (2010), Garratt et al. (2014) and Rossi and Sekhposyan (2014) consider closely related density forecasting exercises. We compare the combination forecasts resulting from both the LOP and our proposed Empirically-transformed LOP (EtLOP) using an evaluation sample from 1990:1 to 2017:2. Relative to the more conventional LOP, the EtLOP improves forecasting performance by 8-9% in terms of both point and density forecasting metrics, namely, the Root Mean Squared Forecast Error (RMSFE) and the Continuous Ranked Probability Score (CRPS).

These findings have particular importance for monetary policymakers concerned with interpreting the SPF and the short-term forecasts from central banks based on density combinations, such as the System for Averaging Models (SAM) developed at Norges Bank. Despite the scope for non-Gaussian macroeconomic variables, many private sector forecasters and central bank forecasting systems find models that are (approximately) Gaussian and linear easy to interpret. Our computationally-convenient EtLOP approach provides a practical route to transform combination density forecasts for a non-Gaussian macroeconomic target variable.

Our approach builds on empirical copula papers by, among others, Deheuvals (1979), Deheuvals (1981), Velásquez-Giraldo et al. (2018) and Coe and

Vahey (2020) by fitting ECDFs with non-parametric methods. Recent macroeconomic applications with semi-parametric copulas utilising non-parametrically fitted ECDFs include Smith and Vahey (2016), Karagedikli et al. (2019), Amengual et al. (2020) and Loaiza-Maya and Smith (2020). Odendahl (2018) uses a parametric copula to model the dependence in the SPF.

In contrast, our approach does not involve fitting a copula model; we do not estimate the dependence in the combination. Our methodology is effective for combinations with a large number of experts relative to the length of the time series available for the target variable—circumstances which blight dependence estimation and where opinion pooling is particularly appealing.

Kascha and Ravazzolo (2010) compare the properties and the forecast performance of the LOP and the geometric averaging Logarithmic Opinion Pool (LogOP). Aastveit et al. (2019) and Rossi (2019) provide discussions of various weighting schemes and recent developments in the opinion (prediction) pooling literature.

Whereas we focus on non-parametric methods to transform the distribution of the LOP forecasts to match that of the target variable, Ranjan and Gneiting (2010) and Gneiting and Ranjan (2013) propose beta transforms. Bassetti et al. (2018) argue for mixtures of betas. In contrast, Ganics (2017) proposes a re-weighting algorithm to match the desired distribution.

A distinct literature addresses the use of external information to improve the accuracy of density forecasts. For example, Galvao et al. (2020) use entropic tilting for density combinations with U.K. survey information. Allayioti (2020) considers tilting BVAR predictions using empirically-transformed SPF forecasts.

The remainder of this paper is structured as follows. In Section 2, we set out our methodology for empirically-transformed opinion pools. In Section 3,

we apply our methodology to forecast densities for U.S. inflation. We present our results in Section 4, and in the final section, we draw some conclusions.

2 A Framework for Opinion Pooling

In this section, we present the details of our proposal to empirically transform the predictive densities from the LOP for a single target variable. We describe briefly conventional opinion pooling, contrasting with our own approach, and then discuss some practical considerations.

2.1 Conventional Opinion Pooling

We begin by describing the LOP and the LogOP. Kascha and Ravazzolo (2010) provide a discussion of both.

In the opinion pooling framework, aggregation is performed by a "decision maker" with no knowledge of how the individual experts make predictions. The decision maker combines out-of-sample forecasts for the target variable from the many experts (models) to give the LOP combination density forecast:

$$p^{LOP}(\pi_\tau) = \sum_{j=1}^{J} w_{j,\tau}\; g(\pi_\tau \mid I_{j,\tau}), \qquad \tau = \underline{\tau}, \ldots, \overline{\tau}, \tag{1}$$

where $g(\pi_\tau \mid I_{j,\tau})$ are the (in this example) one step ahead density forecasts from expert (model) j, $j = 1, \ldots, J$, for the target variable π_τ (inflation in our application), conditional on the information set $I_{j,\tau}$. The publication delay in the production of real-time macroeconomic data ensures that this information set contains lagged variables, here assumed to be dated $\tau - 1$ and earlier. The non-negative weights, $w_{j,\tau}$, in this finite mixture sum to unity and potentially change through time in the evaluation period $\tau = \underline{\tau}, \ldots, \overline{\tau}$; see the discussion in, for example, Garratt et al. (2014). For simplicity, we utilise equal weights

in what follows. Multi-step forecasting (by either direct or iterative methods) and time-varying weights represent well-known extensions.

In contrast to the LOP, the LogOP combination density forecast is:

$$p^{LogOP}(\pi_\tau) = \frac{\prod_{j=1}^{J} g(\pi_\tau \mid I_{j,\tau})^{w_{j,\tau}}}{\int \prod_{j=1}^{J} g(\pi_\tau \mid I_{j,\tau})^{w_{j,\tau}} d\pi_\tau}, \quad \tau = \underline{\tau}, \ldots, \overline{\tau}, \tag{2}$$

where the constant denominator ensures a proper combination density forecast. The LogOP is linear in its logarithmic form.

To fix ideas, following an example provided by Kascha and Ravazzolo (2010), we consider a one-period combination by the decision maker for a single observation of the target variable based on the predictive densities supplied by two Gaussian experts. Both panels of Figure 1 plot the experts' forecasts, where the prediction of Expert 1 has mean -2.0 and standard deviation 1 and that of Expert 2 has mean 2.0 and standard deviation 2.0. The top panel also plots the LOP forecast (assuming equal weights) which is bimodal, with a slightly higher peak associated with the forecast mean of Expert 1. The bottom panel plots the LogOP forecast, which is unimodal, but where the central mass sits between the twin peaks of the LOP (shown in the top panel).

This simple example illustrates several relevant features of conventional opinion pooling with Gaussian experts. First, in the case of LOP, although the experts' density forecasts are individually Gaussian, the combined LOP density is non-Gaussian. Second, the LOP often puts more mass in the tails than LogOP. Third, the LOP tends to preserve disagreement about the central probability mass of the experts, whereas LogOP tends to consolidate. Regardless of the type of prediction pool, this example shows that the aggregated

forecast does not always inherit the distribution of the experts' predictions. Moreover, in the Kascha and Ravazzolo (2010) example illustrated in Figure 1 with equal weights, the distribution of the conditional density forecast is unrelated to the sample history of the target macroeconomic variable.

2.2 Empirically-transformed Pooling

Our methodology involves reshaping the combination density forecasts from the LOP (and the LogOP) using the unconditional distribution of the target variable. The ECDF is the distribution function for the given sample—a step function that represents the entire history of the observations for that variable. The ECDF is, by construction, well-calibrated; whereas, as we noted previously, the LOP combination forecast density is not generally well-calibrated even if the individual forecasts being combined are well calibrated. Among others, Rosenblatt (1952), Diebold et al. (1998), Galbraith and van Norden (2012) and Rossi and Sekhposyan (2019) discuss calibration and the properties of Probability Integrals Transforms (PITs).

Our algorithm reshapes the LOP combination density forecast by adapting methods developed for pseudo-random number generation. The Smirnov transform allows a researcher to generate a conditional density forecast with the same distribution as a (stationary) target variable via the inverse ECDF. The approach is often used to generate pseudo-random numbers from a known but non-parametric distribution. The empirical copula literature adapts the same approach for prediction from a non-parametric copula density with non-parametric marginal distributions. Generating random numbers with a known parametric distribution simply involves substituting a parametric CDF for the ECDF. Because the inverse of the ECDF is used, the methodology is sometimes referred to as inverse transform sampling. In our algorithm, the Smirnov

transform provides a convenient way to convert probabilities to the scale of the target variable. The LOP density forecasts could be easily converted to the unit interval with a known distribution for the experts' forecasts. Since this distribution is unknown, we use a proxy based on pooling the many experts' density forecasts across time.

For expositional ease, we describe our algorithm for a two-experts, single forecast horizon case, considering a single LOP combination forecast density for the target variable. (Our forecasting U.S. inflation application that follows extends consideration to many experts, with multiple forecast origins and horizons in a real-time expanding window forecast evaluation.) The algorithm is broken into four steps.

1. Construct a sample proxy distribution for the conventional combination density forecast by pooling the history of individual forecast densities, fitted with non-parametric methods. Denote this proxy distribution, $\phi(LOP_t)$.

2. Convert draws from the LOP combination density forecast to quantiles using $\phi(\cdot)$. In effect, the LOP forecast density swarm is mapped into a "pseudo swarm" on the unit interval.

3. Fit the ECDF of the target time series variable, π_t, using non-parametric methods denoted $F(\pi_t)$.

4. Convert draws from the pseudo swarm to the scale of π_t using the inverse ECDF, $F^{-1}(\cdot)$, the quantile function.

We emphasise that, as with the LOP (and the LogOP), our approach does not estimate the dependence structure between experts. Under the information assumptions of conventional linear opinion pooling, the decision maker

assumes that the experts' information sets are conditionally independent; see, for example, the discussion in DeGroot and Mortera (1991). In contrast to our approach, a copula combination methodology would exploit Sklar's Theorem to separate dependence from the margins and fit both the dependence and the marginal distributions. Karagedikli et al. (2019) provide examples based on copula combinations of point forecasts.

We illustrate our algorithm by considering the example in Kascha and Ravazollo (2010) discussed previously. Figure 2 plots the same individual experts' forecast densities as in Figure 1, but also displays the empirically-transformed linear (upper panel) and logarithmic combinations (lower panel), EtLOP and EtLogOP, respectively. For this illustration, we used the same sample for U.S. inflation as in our application below. [1]

Looking at Figure 2, in the case of the linear combination (upper panel), the EtLOP density forecast preserves the uni-modality of the individual (Gaussian) densities, with the peak closer to the forecast mean of Expert 1, with a long right tail. This contrasts with the conventional LOP combination shown in Figure 1 (top panel), which is bi-modal.

Turning to the empirically-transformed LogOP (lower panel), the combination density forecast is again single peaked. Relative to the EtLOP case (upper panel), the empirically-transformed LogOP, EtLogOP, is more peaked. The EtLogOP modal forecast differs little from EtLOP, with a similar long right tail. Comparing the EtLOP combination forecast to the conventional LogOP displayed in Figure 1 (lower panel), the conventional forecast density is more diffuse and symmetric.

[1] Figure 4 plots the PDF corresponding to the fitted ECDF. We used the kernel smoother `ksdensity` from MATLAB to compute the ECDF for the inflation sample. We treated the bandwidth as an unknown parameter and in our application below check the robustness over a specified interval. Our illustrative example uses the same bandwidth value as reported in our main results.

3 Application: Forecasting U.S. Inflation

To demonstrate the predictive effectiveness of our approach, we apply the algorithm to conventional combination density forecasts for quarterly U.S. inflation using the target dates from 1990:1 to 2017:2 based on a model space similar to Garratt et al. (2011).

3.1 Expert Space, Forecasts and Data Considerations

Each expert utilises a (unique) bivariate VAR model space for inflation, π_t, and the output gap (the deviation of real output from potential), y_t. The standard theory that aggregate demand, captured by the output gap, influences the movements in inflation (with unknown time lags), provides some foundation for the empirical specification, allowing for simultaneity. The j^{th} VAR model takes the form:

$$\begin{bmatrix} \pi_t \\ y_t^j \end{bmatrix} = \begin{bmatrix} a_{\pi\pi}^j & a_{\pi y}^j \\ a_{y\pi}^j & a_{yy}^j \end{bmatrix} \begin{bmatrix} \pi_{t-1} \\ y_{t-1}^j \end{bmatrix} + \begin{bmatrix} \epsilon_{\pi t}^j \\ \epsilon_{yt}^j \end{bmatrix}, \qquad t = 1, \ldots, T, \quad (3)$$

where $[\epsilon_{\pi t}^j, \epsilon_{yt}^j]' \sim i.i.d.\ N(\mathbf{0}, \mathbf{\Sigma}^j)$. That is, we consider a baseline VAR specification in which the output gap measure has been varied to give J linear and Gaussian VAR models, indexed $j = 1, \ldots, J$. For expositional ease, we ignore the intercept and restrict the lag order of the J VARs to one. Following Garratt et al. (2011), our VAR model space uses seven output gap measures derived from the set of univariate off-model filters considered by Orphanides and van Norden (2002, 2005).

We define the output gap as the difference between observed output and unobserved potential (or the trend component of) output. We denote the (logarithm of) actual output in t as q_t, and let μ_t^j be its trend using definition

j, where $j = 1, \ldots, J$. The output gap, y_t^j, is therefore defined as the difference between actual output and its j^{th} trend measure at time t. We assume the following linear trend-cycle decomposition:

$$q_t \;=\; \mu_t^j + y_t^j. \tag{4}$$

The seven methods of univariate trend extraction in our VAR model space are: quadratic, Hodrick-Prescott (HP), forecast-augmented HP, Christiano and Fitzgerald, Baxter-King, Beveridge-Nelson and Unobserved Components. We summarize these seven well-known univariate filters in Appendix 1.

In our application, we vary a single auxiliary assumption to generate the model space. Specifically, we vary the lag length in the VAR.[2] If we have J output gap measures, and for any given y_t^j we have L variants defined by different values of the maximum lag length, then in total we have $J \times L$ models, each with a corresponding forecast of inflation (and the output gap) from the VAR model space. We restrict L to a maximum of four and therefore we consider 7×4 models—28 forecasts from the experts to be combined.

Although the motivation for deploying these models stems from their common usage by central banks around the world, Orphanides and van Norden (2005) note indifferent real-time out-of-sample forecasting performance. In contrast, Garratt et al. (2011) find improved performance for ensembles of VARs in terms of real-time point and density forecast accuracy with density combinations.

3.2 Data

Orphanides and van Norden (2002, 2005) stress that output gap measures are subject to considerable data revisions. Failing to account for this by us-

[2]For ease of exposition, we fixed this at one in equation (3).

ing heavily-revised data can mask real-time predictive content. Since we are interested in real-time prediction, parameter estimation is recursive for all specifications. Each recursion uses a different vintage of data, where a vintage of data is the vector of time series observations available from a data agency at the forecast origin.

The quarterly real-time real gross domestic product (GDP) U.S. dataset has 112 vintages, with the first vintage dated 1990:1 and the last 2017:4. The raw data for GDP (in practice, Gross National Product, GNP, for some vintages) are taken from the Federal Reserve Bank of Philadelphia's Real-Time Data Set for Macroeconomists. The data comprise successive vintages from the National Income and Product Accounts, with each vintage reflecting the information available around the middle of the respective quarter. Croushore and Stark (2001) provide a description of the real-time GDP database. The GDP deflator price series used to measure inflation is constructed analogously. We define inflation (output growth) as the first difference in the logarithm of the GDP deflator (GDP) multiplied by 400.

Figure 3 displays inflation (upper panel) and real output growth (lower panel) for our evaluation period, from 1990:1 to 2017:2 based on the final vintage of data.[3] For the observations displayed prior to the Great Recession—the Great Moderation—inflation exhibits lower volatility and a higher unconditional mean. A striking feature of the Great Recession and its aftermath is the increased threat of low inflation.

In contrast, during the run up to the Great Recession, between 2003 and 2006, there are several realisations of high inflation. The upward spikes apparent in various inflation measures for the U.S. during this period are often regarded as (the response to) relative price movements, and, in particular,

[3]The empirical analysis which follows uses a time sequence of vintages.

commodity prices. See, for example, the analysis of Garratt and Petrella (2019).

3.3 Forecast Combination and Empirical Transformation

The decision maker recursively combines the forecast densities from the experts. Each expert uses an expanding window for parameter estimation. For the first recursion the estimation sample is 1970:1 to 1989:4 (window size 80 observations) and the last 1970:1 to 2017:1 (window size 189 observations).

As our U.S. GDP deflator data are released with a one quarter lag, the first vintage, dated 1990:1, contains time series observations from 1970:1 to 1989:4, and the last vintage, dated 2017:4, has data from 1970:1 to 2017:3. Following Clark and McCracken (2010) and others, we use the second estimate as the target "final" data. For example, when evaluating the $h = 1$ forecast (nowcast) for 2017:2, we use the 2017:4 vintage observation of inflation for 2017:2.

For each VAR (expert), we estimate the parameters using Ordinary Least Squares over an expanding window. It is straightforward to produce forecast densities for both inflation and the (j^{th}) output gap through our evaluation period: $\tau = \underline{\tau}, \ldots, \overline{\tau}$ where $\underline{\tau}$ =1990:1 and $\overline{\tau}$ =2017:2 (110 quarterly observations). Our experts focus on forecasting inflation, given the output gap data as well as lagged inflation data. [4]

To deploy our empirical transformation algorithm, the decision maker must fit a distribution for inflation, π_t. The (smoothed) probability density function corresponding to the fitted ECDF plotted in Figure 4 uses the final vintage of inflation data.[5] The distribution displays some asymmetry, with the right tail extending relatively far from the central mass. The null hypothesis of

[4] Recall that we treat the "true" output gap as latent.

[5] Since the decision maker uses recursive fitting based on expanding windows of data, the fitted distributions for inflation vary slightly by forecast origin in practice.

normality is rejected at the 1% significance level for all vintages.

In the following section, we compare and contrast the forecast performance of our candidate density combinations, the LOP (LogOP) and the EtLOP (EtLogOP), to assess the relative accuracy of the empirically transformed opinion variants. The performance metrics gauge point forecast accuracy, using the Root Mean Squared Forecast Error (RMSFE), and density forecast accuracy, using the average Continuous Ranked Probability Score (CRPS) proposed by Gneiting and Raftery (2007).

4 Results

In this section, we report results for the one-step ahead forecasts from the forecast origins. Appendix 2 and Figures 9 to 12 provide results for the (iterative) four-step ahead case which display quantitatively similar results. All results reported refer to the equal weighted LOP (and the EtLOP) combinations.[6]

4.1 Bandwidth Selection

Table 1a presents out-of-sample forecast metrics for EtLOP for a variety of values of the bandwidth for $F(\cdot)$, denoted bw. Table 1b provides similar information for EtLogOP. The first column reports how the Brier score (Brier, 1950) varies with bw, for the event defined as inflation below its unconditional mean:

$$BS \;=\; (1/M)\sum_{\tau=\underline{\tau}}^{\overline{\tau}}(o_\tau - f_\tau)^2, \tag{5}$$

where M is the number of periods in the out-of-sample evaluation, $\overline{\tau}$ minus $\underline{\tau}$, o_τ is the outcome (1 if inflation below mean, 0 otherwise), and f_τ is the real-time probability that inflation is below its mean (generated by the candidate

[6]Recursive weighted combinations based on the time-averaged logarithmic score or the time-averaged CRPS weights give similar results.

opinion pools). Low Brier scores are preferred. The second column reports the Root Mean Squared Forecast Error (RMSFE) as a ratio to the benchmark LOP (LogOP). The third column displays the (time-averaged) logarithmic score and the fourth column the (time-averaged) CRPS.

The four columns of Table 1a (Table 1b) suggest a relatively flat response over bw for the Brier score, RMSFE, the logarithmic score and the CRPS for the EtLOP. Given the fairly robust performance across bandwidths, we present results below for a representative case, where $bw = 0.975$, with the corresponding rows in Table 1 in bold for emphasis.

4.2 Comparing Density Combinations

The four rows of Table 2 display the main results of the paper, with the principle concern being the contrast between EtLOP and the benchmark LOP, with EtLogOP and LogOP provided for context. The EtLOP (EtLogOP) results refer to the $bw = 0.975$, forecast horizon $h = 1$ case. The first column of Table 2 reports the RMSFE and the second column reports the time-averaged CRPS over the evaluation sample. Columns 3 to 5 give the tail-weighted, right tail-weighted, and left tail-weighted CRPS, respectively. Except for the RMSFE and CRPS statistics for the LOP model (first row, in italics), which are reported as absolute values, all other values in columns 1 to 5 are computed as ratios to the LOP benchmark. Ratios less than one, for both RMSFE and CRPS, indicate an improvement in forecast performance, relative to the LOP benchmark model.

The RMSFEs reported in the first column indicate a gain for both the EtLOP (and the EtLogOP) of approximately 8% over the LOP benchmark. The CRPS values reported in column 2 likewise indicate a performance gain for EtLOP (and EtLogOP) of around 9% over the LOP benchmark. This

performance gain is slightly greater when considering the tail-weighted CRPS, either using both tails or just the right tail (columns 3 and 4), whereas the left tail-weighted CRPS has a slightly smaller performance gain (column 5) from the empirical-transformation.

To summarise the results so far, empirically transformed opinion pools outperform their more conventional counterparts in terms of point and density forecasting performance. To give further context, a Bayesian estimated univariate unobserved components stochastic volatility (UCSV) model delivered forecast performance comparable to the empirically transformed opinion pools.[7]

The upper and lower panels of Figure 5 show the 5% and 95% uncertainty bands of the forecast densities, as well as the mean forecasts, for inflation for the LOP and the EtLOP, respectively. The mean forecasts of the EtLOP model are quite close to those of the LOP.

There are differences in the shape of the density forecasts. For example, in 2010 the 5% threshold of the density forecast band in the LOP extends below -1%, whereas in the EtLOP it is around -0.5%. A similar pattern can be seen elsewhere, for example in 2002 and in the second half of 2015. This difference reflects the EtLOP's ability to match the skewness of the sample inflation data. Furthermore, a careful comparison of the two panels (e.g. along the vertical lines) reveals that the EtLOP's 90% band is narrower than the LOP's 90% band.

The upper and lower panels of Figure 6 display similar characteristics when comparing the LogOP and EtLogOP combination density forecasts, with the empirically-transformed combination displaying more skew and being some-

[7] Appendix 3 describes the UCSV model of Chan and Song (2018). We experimented with prediction pools of UCSV experts but found no further improvement in forecast performance. Carriero et al. (2020) explore the ability of SV specifications to produce asymmetric density forecasts.

what sharper than the conventional counterpart.

Figure 7 displays the density forecast resulting from the four specifications for the target observation 2009:2, when inflation was unusually low. The densities of both empirically-transformed opinion pools have less probability mass on high inflation realisations and are somewhat less diffuse than their conventional counterparts.

Figure 8 provides a visual confirmation of the performance differentials in terms of the levels of RMSFE and CRPS computed recursively through the evaluation sample. The EtLOP (EtLogOP) dominates the LOP (LogOP) throughout, apart from the first year of the evaluation sample.

5 Conclusions

In this paper, we have proposed a methodology to improve the accuracy of opinion pools. Our approach involves transforming the conventional combination density forecasts using an ECDF to match the unconditional distribution of the sample data. In our U.S. inflation application, using an evaluation sample considering the Great Recession, we combined density forecasts from a system of VAR models. We demonstrated that the Empirically-transformed LOP results in considerably improved forecast performance relative to the more conventional opinion pool.

Our approach provides a computationally-convenient methodology to handle a non-Gaussian distributed macroeconomic target variable within the opinion pooling framework. The approach is effective in combinations with large numbers of experts relative to the length of the time series.

Further work in this area should focus on applying the methodology to survey forecasts and investigating the scope for improving the accuracy of joint forecast distributions.

Figure 1: Conventional Combinations and Expert Density Forecasts

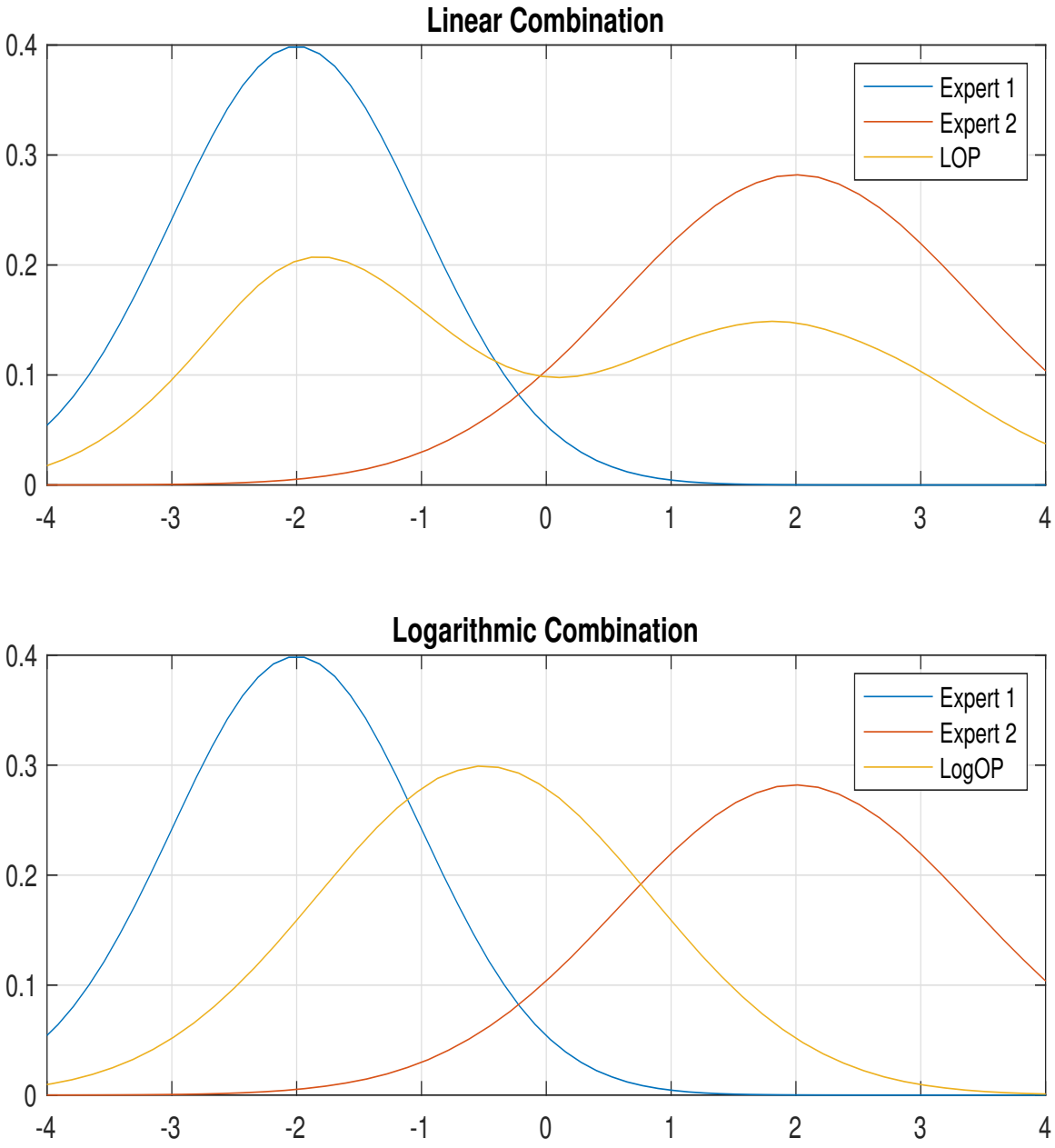


Figure 2: Empirically-transformed Combination and Expert Density Forecasts

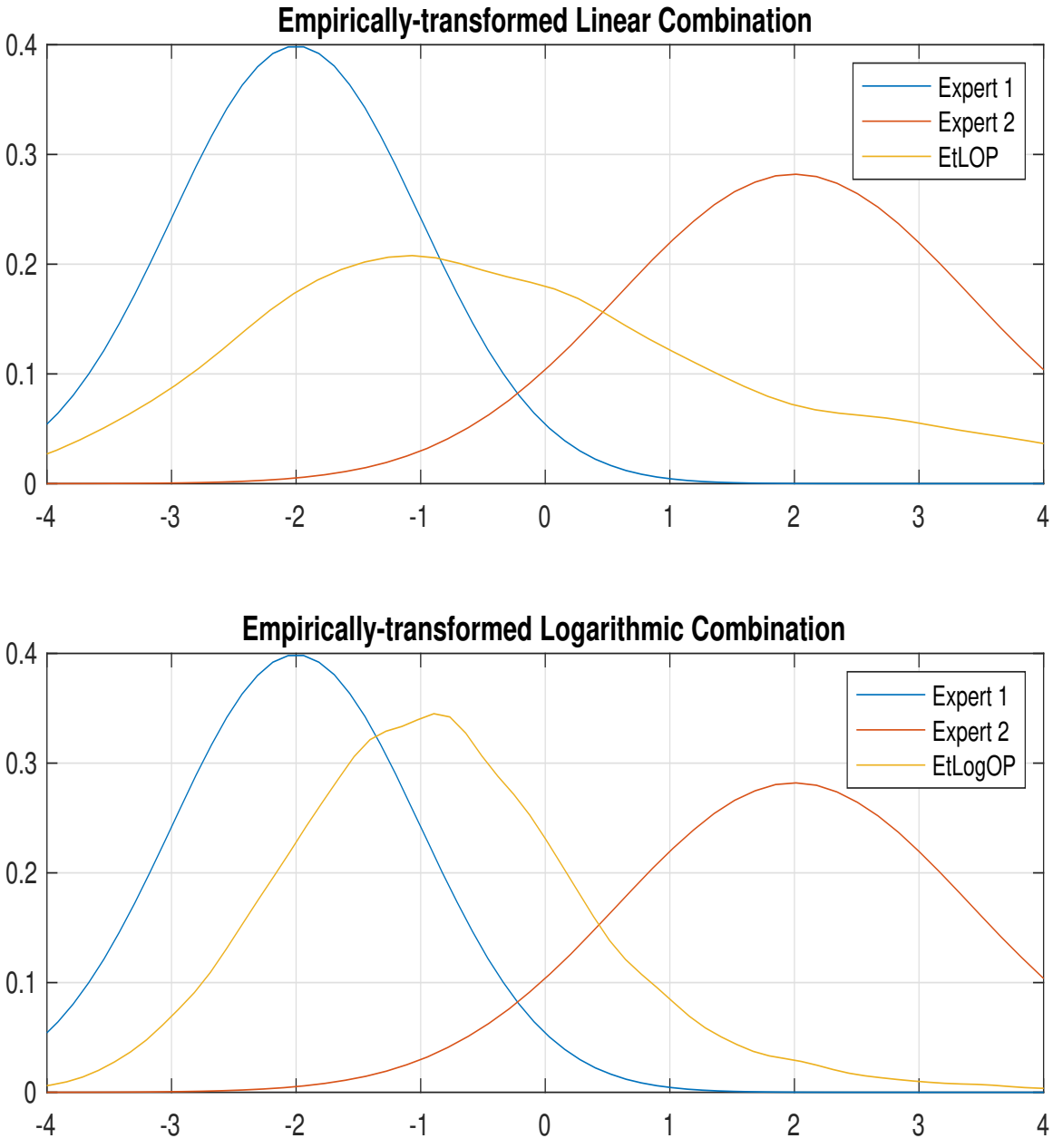


Figure 3: U.S. Inflation and Real Output Growth

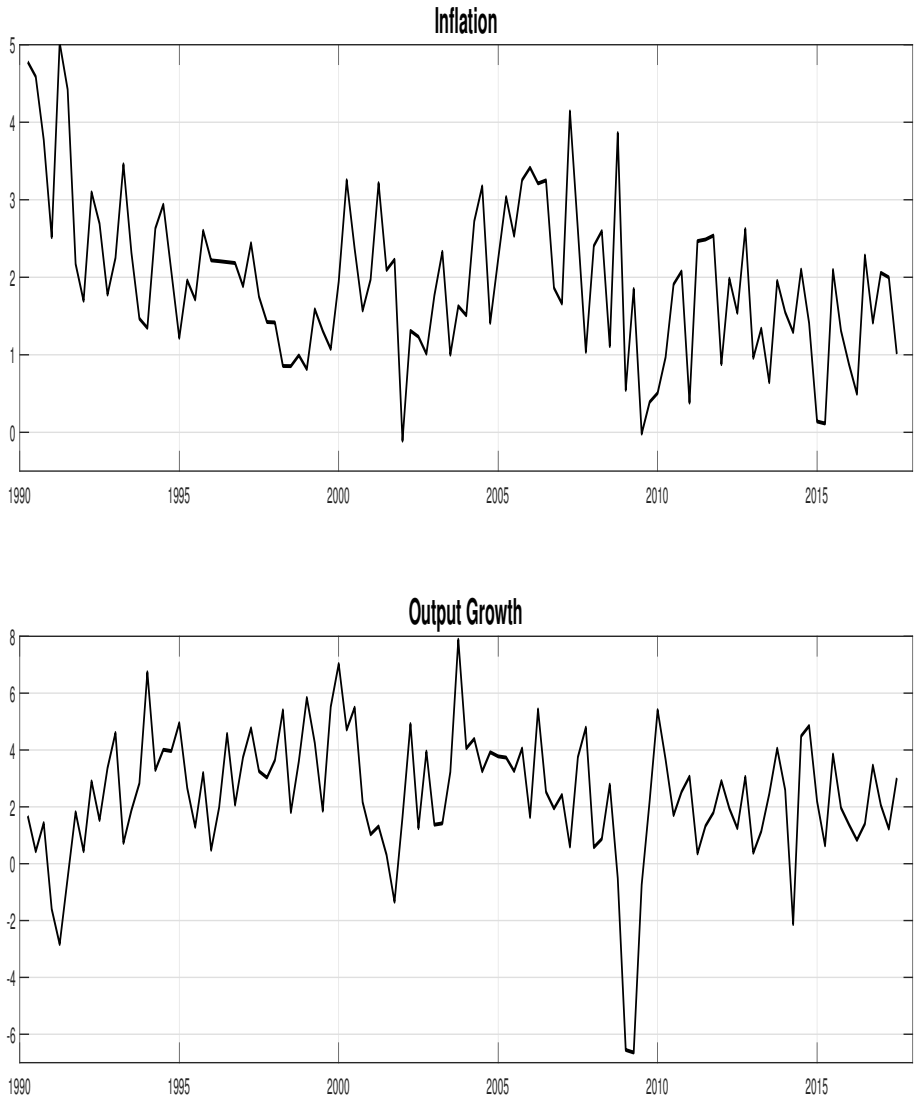


Figure 4: Empirical Density for U.S. Inflation

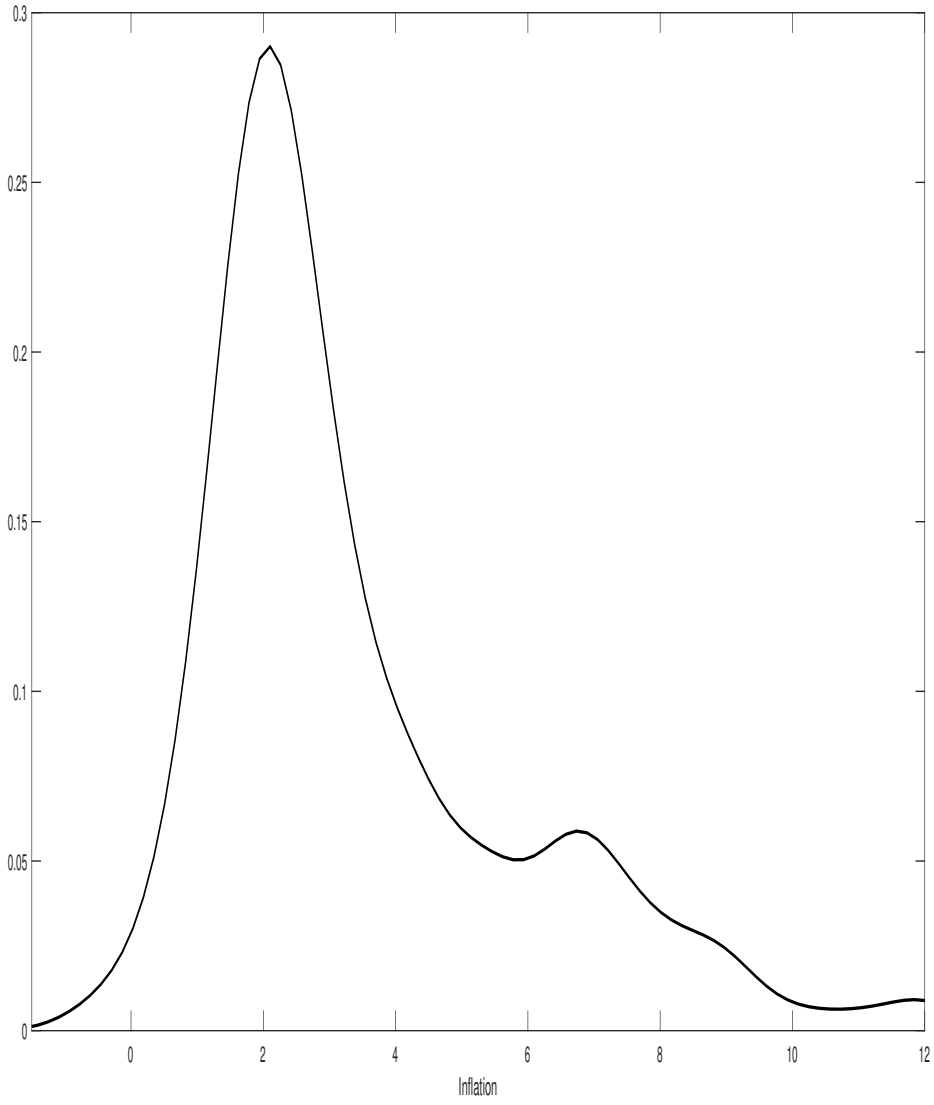


Figure 5: LOP and EtLOP Forecasts

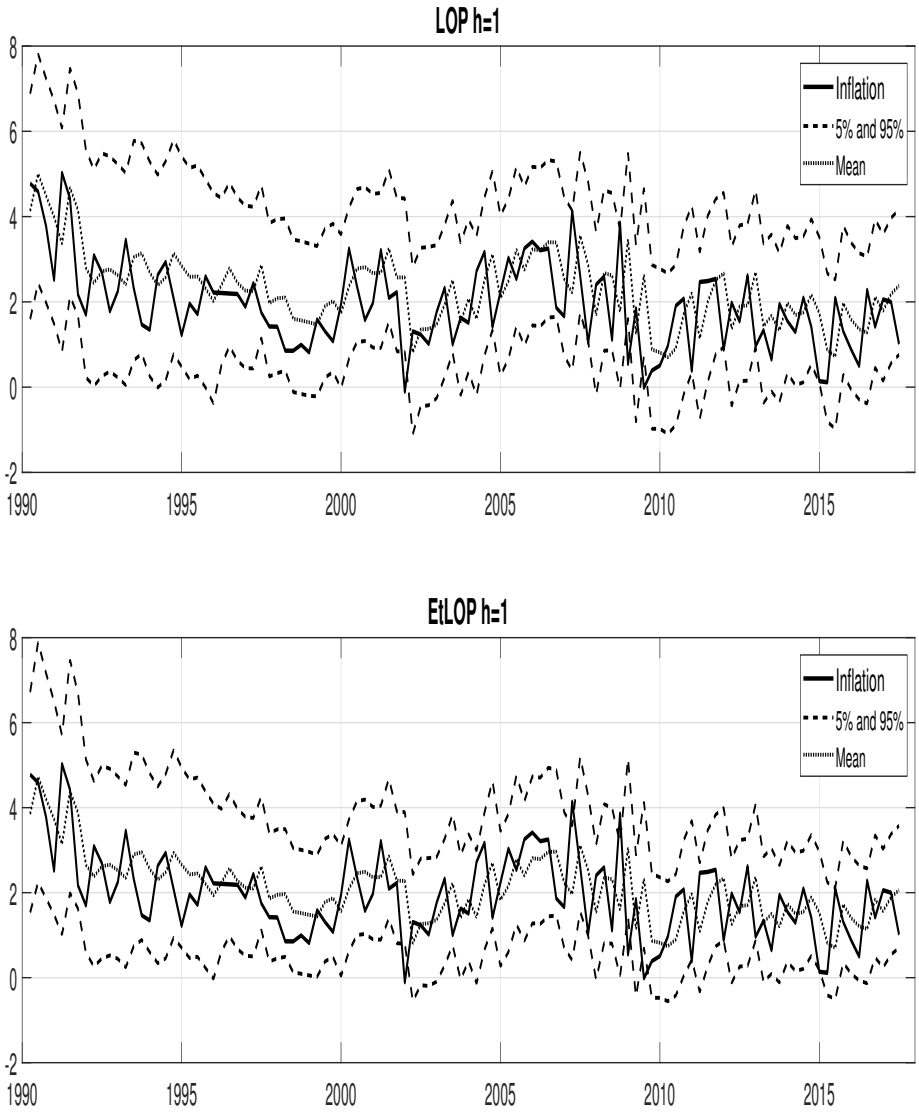


Figure 6: LogOP and EtLogOP Forecasts

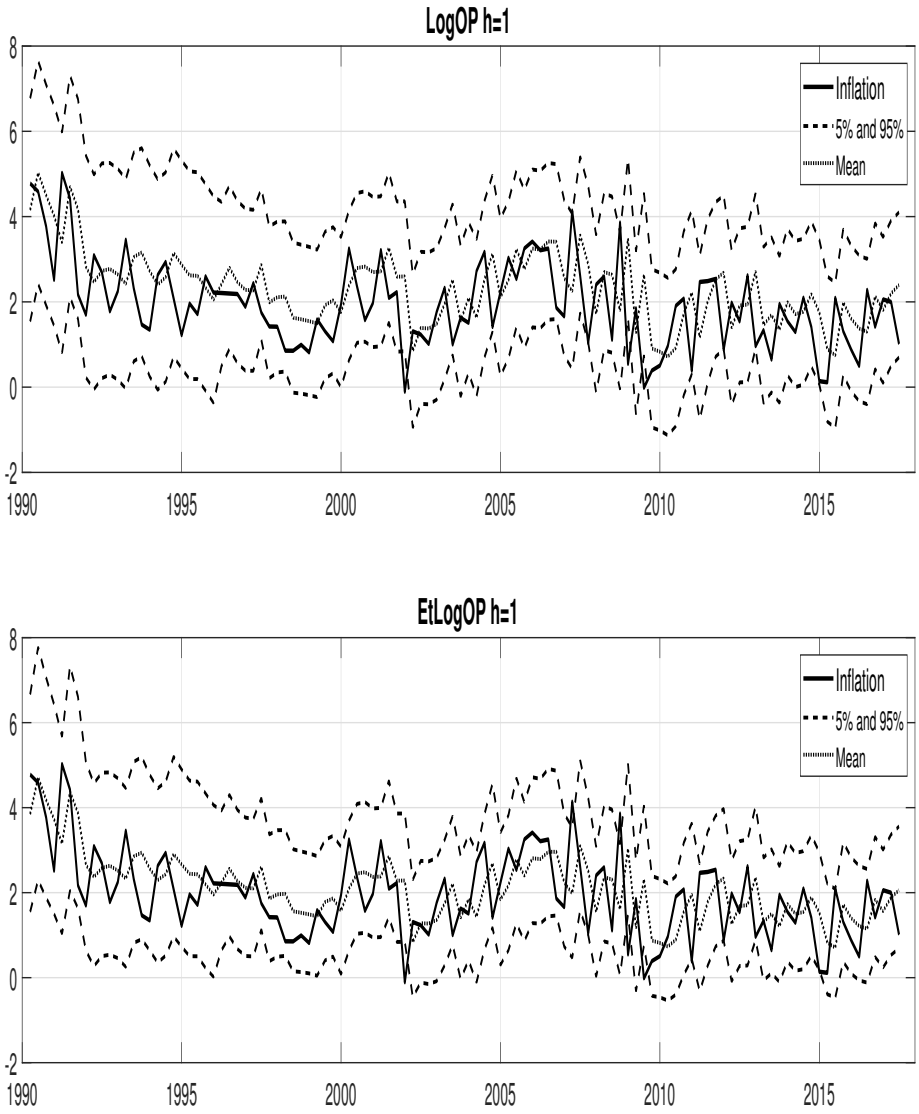


Figure 7: Forecast Densities for 2009:02

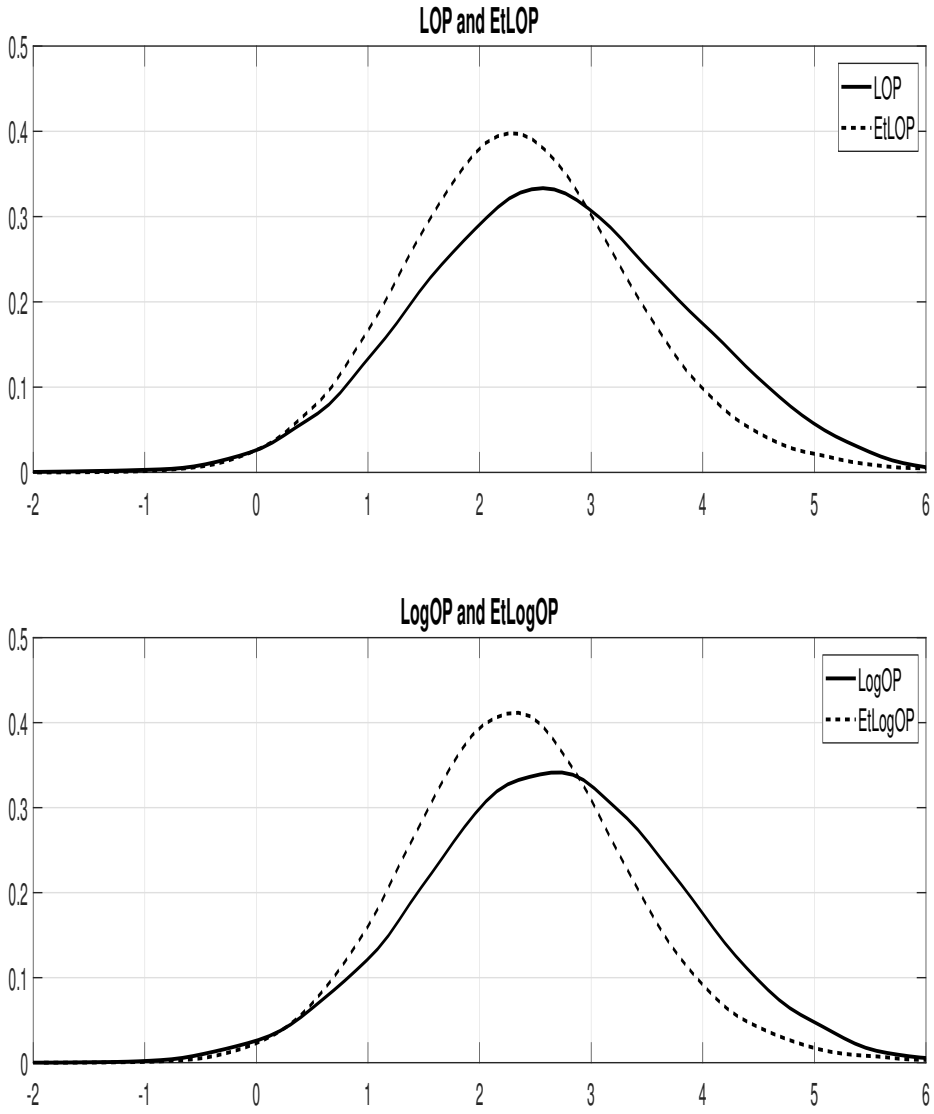


Figure 8: Recursive RMSFE and CRPS

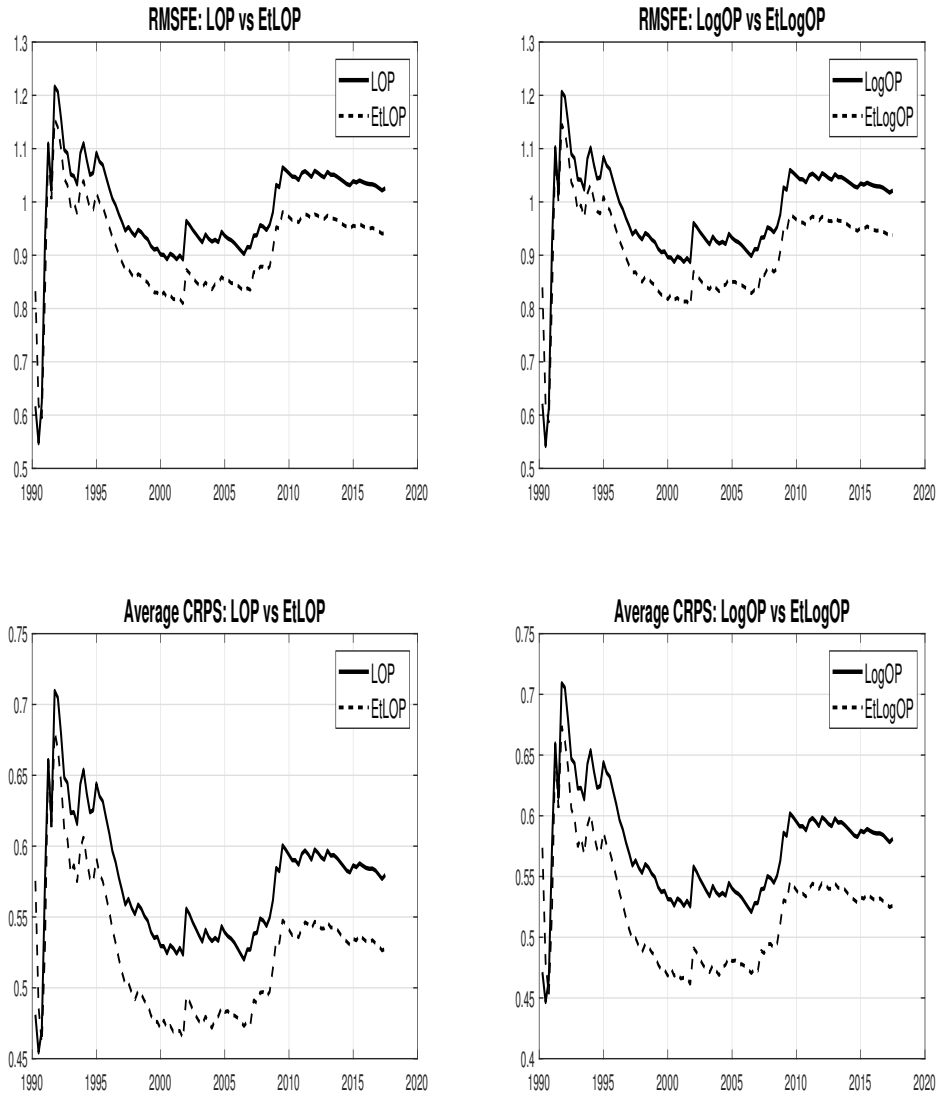


Table 1: Performance of Empirically-transformed Combinations, by Bandwidth (bw)

(a) EtLOP

bw	BS	RMSFE Ratio	LS	CRPS
0.800	0.927	0.834	-3.757	0.530
0.825	0.923	0.836	-3.746	0.529
0.850	0.917	0.832	-3.737	0.529
0.875	0.919	0.831	-3.729	0.528
0.900	0.916	0.830	-3.722	0.528
0.925	0.908	0.830	-3.715	0.528
0.950	0.912	0.829	-3.709	0.528
0.975	**0.906**	**0.829**	**-3.704**	**0.528**
1.000	0.907	0.829	-3.699	0.528
1.025	0.908	0.829	-3.695	0.529
1.050	0.905	0.829	-3.691	0.529
1.075	0.898	0.829	-3.688	0.530
1.100	0.899	0.829	-3.684	0.530
1.125	0.912	0.830	-3.681	0.531
1.150	0.894	0.830	-3.678	0.532
1.750	0.903	0.831	-3.675	0.533
1.200	0.899	0.831	-3.673	0.531

(b) EtLogOP

bw	BS	RMSFE Ratio	LS	CRPS
0.800	0.926	0.830	-3.717	0.528
0.825	0.921	0.829	-3.705	0.528
0.850	0.917	0.828	-3.696	0.527
0.875	0.925	0.827	-3.687	0.526
0.900	0.917	0.827	-3.679	0.526
0.925	0.913	0.826	-3.672	0.526
0.950	0.911	0.826	-3.666	0.526
0.975	**0.910**	**0.826**	**-3.661**	**0.526**
1.000	0.909	0.825	-3.656	0.526
1.025	0.910	0.825	-3.651	0.527
1.050	0.903	0.826	-3.647	0.527
1.075	0.908	0.826	-3.644	0.528
1.100	0.901	0.826	-3.641	0.528
1.125	0.907	0.826	-3.638	0.529
1.150	0.898	0.827	-3.635	0.530
1.175	0.897	0.827	-3.633	0.531
1.200	0.904	0.828	-3.630	0.531

Notes: The first column reports the scaled Brier Score (BS), for the inflation event of less than the sample mean, calculated by dividing the Brier Score by the measure of data uncertainty, defined as: $\overline{o}(1-\overline{o})$, where $\overline{o}$ is the observed frequency of the event. The second column reports the RMSFE ratio relative to the LOP. Column three reports the log score and column four the CRPS.

Table 2: Forecast Performance of Empirically-transformed and Conventional Combinations, $h = 1$**,** $bw = 0.975$

Model	**RMSFE**	**CRPS**	**CRPS-TW**	**CRPS-RT**	**CRPS-LT**
LOP	*1.026*	*0.580*	*0.129*	*0.184*	*0.171*
EtLOP	0.918**	0.910**	0.899**	0.891**	0.924**
LogOP	0.996**	1.002**	0.992**	0.989**	1.000
EtLogOP	0.914**	0.907**	0.891**	0.886**	0.924**

Notes: The first row in italics reports absolute values of the RMSFE and CRPS statistics for the LOP. The remaining rows report ratios, relative to the LOP, for RMSFE and CRPS. Ratios less than one, for both RMSFE and CRPS, indicate an improvement in forecast performance relative to the LOP. As a rough guide, using a Harvey et al. (1997) small-sample adjustment of the Diebold and Mariano (1995) test, the superscript $**$ denotes significantly different from the LOP at the 5% level for RMSFE. The corresponding statistic for CRPS is denoted similarly. Columns three (CRPS-TW), four (CRPS-RT) and five (CRPS_LT) report tail-weighted, right tail weighted and left tail weighted CRPS; see Gneiting and Ranjan (2011) and Diks et al. (2011).

References

[1] Aastveit, K. A., J. Mitchell, F. Ravazzolo, and H.K van Dijk (2019), "The Evolution of Forecast Density Combinations in Economics", Oxford Research Encyclopedia of Economics and Finance, Oxford University Press.

[2] Adrian, T., N. Boyarchenko and D. Giannone (2019), "Vulnerable Growth", *American Economic Review*, 109, 4, 1263-1289.

[3] Allayioti, A. (2020), "Bayesian Entropic Tilting for Macroeconomic Variables with Reshaped Survey Information", mimeo, Warwick Business School, University of Warwick.

[4] Amengual, D., E. Sentana and Z. Tian (2020), "Gaussian Rank Correlation and Regression", CEPR Discussion Papers 14914, June.

[5] Bassetti, F., R. Casarin and F. Ravazzolo (2018), "Bayesian Nonparametric Calibration and Combination of Predictive Distributions", *Journal of the American Statistical Association*, 113, 552, 675-685.

[6] Baxter, M, and R.G. King (1999), "Measuring Business Cycles: Approximate Band-Pass Filters for Economic Time Series", *Review of Economics and Statistics*, 81, 594-607.

[7] Beveridge, S., and C.R. Nelson (1981), "A New Approach to Decomposition of Time Series into Permanent and Transitory Components with Particular Attention to Measurements of the Business Cycle", *Journal of Monetary Economics*, 7, 151-174.

[8] Brier, G.W. (1950), "Verification of Forecasts Expressed in Terms of Probability", *Monthly Weather Review*, 78, 1-3.

[9] Chan, J.C.C., and Y. Song (2018), "Measuring Inflation Expectations Uncertainty Using High-Frequency Data", *Journal of Money Credit and Banking*, 50, 1139-1166.

[10] Carriero, A., T.E. Clark and M. Marcellino (2020) "Capturing Macroeconomic Tail Risks with Bayesian Vector Autoregressions", Federal Reserve Bank of Cleveland, Working Paper No. 20-02, January.

[11] Christiano, L. and T.J. Fitzgerald (2003), "The Band Pass Filter", *International Economic Review*, 44, 2, 435-465.

[12] Clark, T.E. and M.W. McCracken (2010), "Averaging Forecasts from VARs with Uncertain Instabilities", *Journal of Applied Econometrics*, January-February, 5-29.

[13] Coe, P.J. and S.P. Vahey (2020), "Financial Conditions and the Risks to Economic Growth in the United States Since 1875", CAMA Working Paper No. 36/2020, Australian National University, April.

[14] Croushore, D. and T. Stark (2001), "A Real-time Data Set for Macroeconomists", *Journal of Econometrics*, 105, 111-130.

[15] DeGroot M.H. and J. Mortera (1991), "Optimal Linear Opinion Pools", *Management Science*, 37, 5, 546-558.

[16] Deheuvels, P. (1979), "La Fonction de Dépendence Empirique et Ses Propriétés. Un Test non Paramétrique d'Indépendence", *Bulletin Royal Belge de l'Académie des Sciences*, 65, 274-292.

[17] Deheuvels, P. (1981), "An Asymptotic Decomposition for Multivariate Distribution-free Tests of Independence", *Journal of Multivariate Analysis*, 11, 1, 102-113.

[18] Diebold, F.X., T.A. Gunther and A.S. Tay (1998), "Evaluating Density Forecasts; with Applications to Financial Risk Management", *International Economic Review,* 39, 863-83.

[19] Diebold, F.X. and R.S. Mariano (1995), "Comparing Predictive Accuracy", *Journal of Business and Economic Statistics*, 13, 253-263.

[20] Diks, C., V. Panchenko, V. and D. van Dijk (2011), "Likelihood-Based Scoring Rules for Comparing Density Forecast in Tails", *Journal of Econometrics*, 163, 2, 215-230.

[21] Galbraith, J.W. and S. van Norden (2012), "Assessing Gross Domestic Product and Inflation Probability Forecasts Derived from Bank of England Fan Charts", *Journal of the Royal Statistical Society*, Series A (Statistics in Society), 175, 3, 713-727.

[22] Galvao, A.B., A. Garratt and J. Mitchell (2020), "Does Judgement Improve Macroeconomic Density Forecasts?", MPF Working Paper 33, Warwick Business School, University of Warwick, May.

[23] Ganics, G. (2017), "Optimal Density Forecast Combinations", Bank of Spain Working Paper No. 1751.

[24] Garratt, A., K. Lee, E. Mise and K. Shields (2008), "Real-Time Representations of the Output Gap", *Review of Economics and Statistics*, 90, 4, 792-804.

[25] Garratt, A., J. Mitchell, S.P. Vahey and E. Wakerly (2011), "Real-time Inflation Forecast Densities from Ensemble Phillips Curves", *North American Journal of Economics and Finance*, 22, 77-87.

[26] Garratt, A., J. Mitchell, and S.P. Vahey (2014), "Measuring Output Gap Nowcast Uncertainty", *International Journal of Forecasting*, 30, 2, 268-279.

[27] Garratt, A., and I. Petrella (2019), "Commodity Prices and Inflation Risk", *MPF Working Paper 23*, Warwick Business School, University of Warwick, May.

[28] Gneiting, T. and A.E. Raftery (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation", *Journal of the American Statistical Association*, 102, 359-378.

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

[30] Gneiting, T. and R. Ranjan (2013), "Combining Predictive Distributions", *Electronic Journal of Statistics*, 7, 7471782.

[31] Harvey, D., S. Leybourne and P. Newbold (1997), "Testing the Equality of Prediction Mean Squared Errors", *International Journal of Forecasting*, 13, 281-291.

[32] Hodrick, R. and E. Prescott (1997), "Post-War U.S. Business Cycles: An Empirical Investigation", *Journal of Money, Banking and Credit*, 29, 1-16.

[33] Jore, A.S., J. Mitchell and S.P. Vahey (2010), "Combining Forecast Densities from VARs with Uncertain Instabilities", *Journal of Applied Econometrics*, 25, 621-634.

[34] Karagedikli, O., S.P. Vahey and E.C. Wakerly (2019), "Improved Methods for Combining Point Forecasts for an Asymmetrically Distributed

Variable", CAMA Working Paper No. 15/2019, Australian National University, February.

[35] Kascha, C. and F. Ravazzolo (2010), "Combining Inflation Density Forecasts", *Journal of Forecasting*, 29, 231-250.

[36] Loaiza-Maya, R. and M.S. Smith (2020), "Real-Time Macroeconomic Forecasting with a Heteroskedastic Inversion Copula", *Journal of Business and Economic Statistics*, 38, 2, 470-486.

[37] Mise, E., T-H. Kim and P. Newbold (2005), "On the Sub-Optimality of the Hodrick-Prescott Filter", *Journal of Macroeconomics*, 27, 1, 53-67.

[38] Odendahl, F. (2018), "Survey-Based Joint Density Forecasts", mimeo, UPF, October.

[39] Orphanides, A. and S. van Norden (2002), "The Unreliability of Output-Gap Estimates in Real Time", *Review of Economics and Statistics*, 84, 4, 569-583.

[40] Orphanides, A. and S. van Norden (2005), "The Reliability of Inflation Forecasts Based on Output-Gap Estimates in Real Time", *Journal of Money Credit and Banking*, 37, 3, 583-601.

[41] Ranjan, R. and T. Gneiting (2010), "Combining Probability Forecasts", *Journal of the Royal Statistical Society Series B*, 72, 71-91.

[42] Rosenblatt, M. (1952), "Remarks on a Multivariate Transformation", *The Annals of Mathematical Statistics*, 23, 470-472.

[43] Rossi, B. (2019) "Forecasting in the Presence of Instabilities: How Do We Know Whether Models Predict Well and How to Improve Them", Barcelona GSE Working Paper Series Working Paper 1161, November.

[44] Rossi, B. and T. Sekhposyan (2014), "Evaluating Predictive Densities of U.S. Output Growth and Inflation in a Large Macroeconomic Data Set", *International Journal of Forecasting*, 30, 3, 662-682.

[45] Rossi, B. and T. Sekhposyan (2019), "Alternative Tests for Correct Specification of Conditional Predictive Densities", *Journal of Econometrics* 208, 2, 638-657.

[46] Smith, M.S. and S.P. Vahey (2016), "Asymmetric Forecast Densities for U.S. Macroeconomic Variables from a Gaussian Copula Model of Cross-Sectional and Serial Dependence", *Journal of Business and Economic Statistics*, 34, 416-434.

[47] Velásquez-Giraldo, M., Canavire-Bacarreza, G., Huynh, K. and D. T. Jacho-Chavez (2018), "Flexible Estimation of Demand Systems: A Copula Approach", *Journal of Applied Econometrics*, 33, 1109-1116.

Appendix 1: Output trend definitions

We summarise the seven univariate detrending specifications below.

1. For the quadratic trend based measure of the output gap we use the residuals from a regression (estimated recursively) of output on a constant and a squared time trend.

2. Following Hodrick and Prescott (1997, HP), we set the smoothing parameter to 1600 for our quarterly U.S. data.[8]

3. Since the HP filter is a two-sided filter it relates the time-t value of the trend to future and past observations. Moving towards the end of a finite sample of data, the HP filter becomes progressively one-sided and its properties deteriorate with the Mean Squared Error (MSE) of the unobserved components increasing and the estimates ceasing to be optimal in a MSE sense. We therefore follow Mise et al. (2005) and mitigate this loss near and at the end of the sample by extending the series with forecasts. At each recursion the HP filter is applied to a forecast-augmented output series (again with smoothing parameter 1600), with forecasts generated from an univariate AR(8) model in output growth (again estimated recursively using the appropriate vintage of data). The implementation of forecast augmentation when constructing real-time output gap measures for the U.S. is discussed at length in Garratt et al. (2008). We deliberately select a high lag order to ensure no important lags are omitted—favouring unbiasedness over efficiency.

4. Christiano and Fitzgerald (2003) propose an optimal finite-sample approximation to the band-pass filter, without explicit modelling of the

[8] We could, of course, allow for uncertainty in the smoothing parameter. We reduce the computational burden in this application by fixing this parameter at 1600.

data. Their approach implicitly assumes that the series is captured reasonably well by a random walk model and that, if there is drift present, this can be proxied by the average growth rate over the sample.

5. We also consider the band-pass filter suggested by Baxter and King (1999). We define the cyclical component to be fluctuations lasting no fewer than six, and no more than thirty-two quarters—the business cycle frequencies indicated by Baxter and King (1999)—and set the truncation parameter (the maximum lag length) at three years. As with the HP filter we augment our sample with AR(8) forecasts to fill in the 'lost' output gap observations at the end of the sample due to truncation.

6. The Beveridge and Nelson (1981) decomposition relies on a priori assumptions about the correlation between permanent and transitory innovations. The approach imposes the restriction that shocks to the transitory component and shocks to the stochastic permanent component have a unit correlation. We assume the ARIMA process for output growth is an AR(8), the same as that used in our forecast augmentation.

7. Finally, our Unobserved Components model assumes q_t is decomposed into trend, cyclical and irregular components

$$q_t = \mu_t^7 + y_t^7 + \xi_t, \quad \xi_t \sim i.i.d.\ N(0, \sigma_\xi^2), \quad t = 1, \ldots, T \tag{A1.1}$$

where the stochastic trend is specified as

$$\mu_t^7 = \mu_{t-1}^7 + \beta_{t-1} + \eta_t, \quad \eta_t \sim i.i.d.\ N(0, \sigma_\eta^2)$$
$$\beta_t = \beta_{t-1} + \zeta_t, \quad \zeta_t \sim i.i.d.\ N(0, \sigma_\zeta^2).$$

Letting $\sigma_\zeta^2 > 0$ but setting $\sigma_\eta^2 = 0$, gives an integrated random walk.

The cyclical component is assumed to follow a stochastic trigonometric process:

$$\begin{bmatrix} y_t^7 \\ y_t^{7*} \end{bmatrix} = \rho \begin{bmatrix} \cos\lambda & \sin\lambda \\ -\sin\lambda & \cos\lambda \end{bmatrix} \begin{bmatrix} y_{t-1}^7 \\ y_{t-1}^{7*} \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix} \tag{A1.2}$$

where λ is the frequency in radians, ρ is a damping factor and κ_t and κ_t^* are two independent white noise Gaussian disturbances with common variance σ_κ^2. We estimate this model by maximum likelihood, exploiting the Kalman filter, and estimates of the trend and cyclical components are obtained using the Kalman smoother.

Appendix 2: Forecast Performance of Empirically-transformed and Conventional Combinations, $h = 4$**,** $bw = 0.975$

Model	**RMSFE**	**CRPS**	**CRPS-TW**	**CRPS-RT**	**CRPS-LT**
LOP	*1.199*	*0.682*	*0.148*	*0.232*	*0.184*
EtLOP	0.894**	0.881**	0.889**	0.871**	0.895**
LogOP	0.980**	0.996**	0.986**	0.986**	0.996**
EtLogOP	0.873**	0.864**	0.867**	0.842**	0.887**

Notes: The first row in italics reports absolute values of the RMSFE and CRPS statistics for the LOP. The remaining rows report ratios, relative to the LOP, for RMSFE and CRPS. Ratios less than one, for both RMSFE and CRPS, indicate an improvement in forecast performance relative to the LOP. As a rough guide, using a Harvey et al. (1997) small-sample adjustment of the Diebold and Mariano (1995) test, the superscript ** denotes significantly different from the LOP at the 5% level for RMSFE. The corresponding statistic for CRPS is denoted similarly. Columns three (CRPS-TW), four (CRPS-RT) and five (CRPS_LT) report tail-weighted, right tail weighted and left tail weighted CRPS; see Gneiting and Ranjan (2011) and Diks et al. (2011).

Figure 9: LOP and EtLOP Forecasts

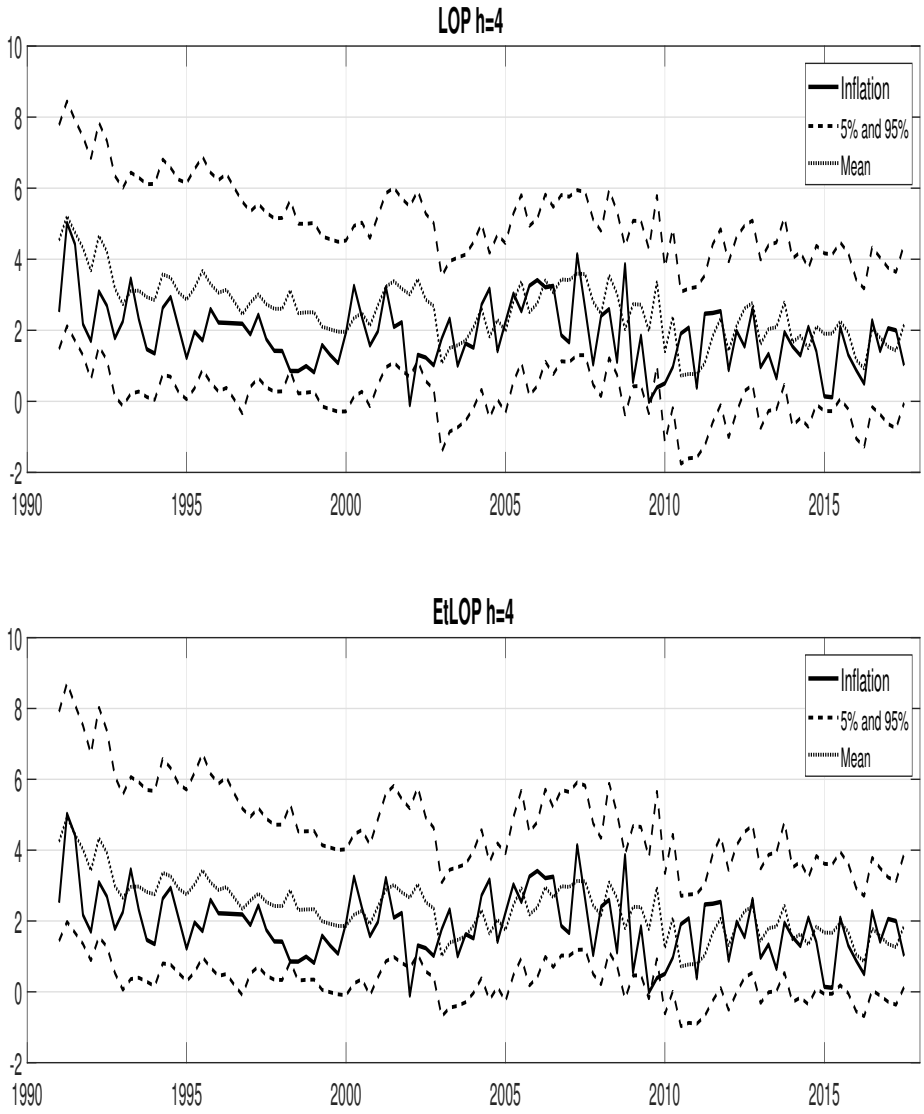


Figure 10: LogOP and EtLogOP Forecasts

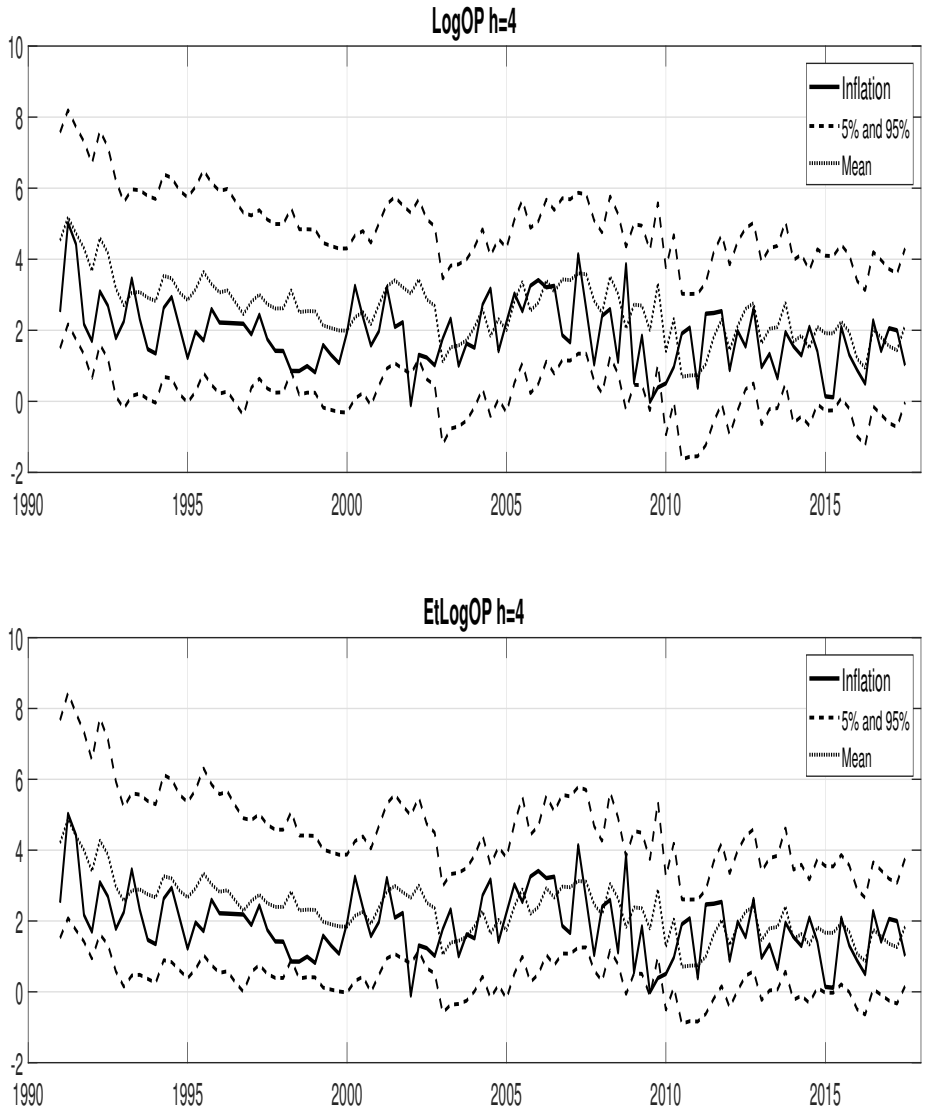


Figure 11: Forecast Densities for 2009:02

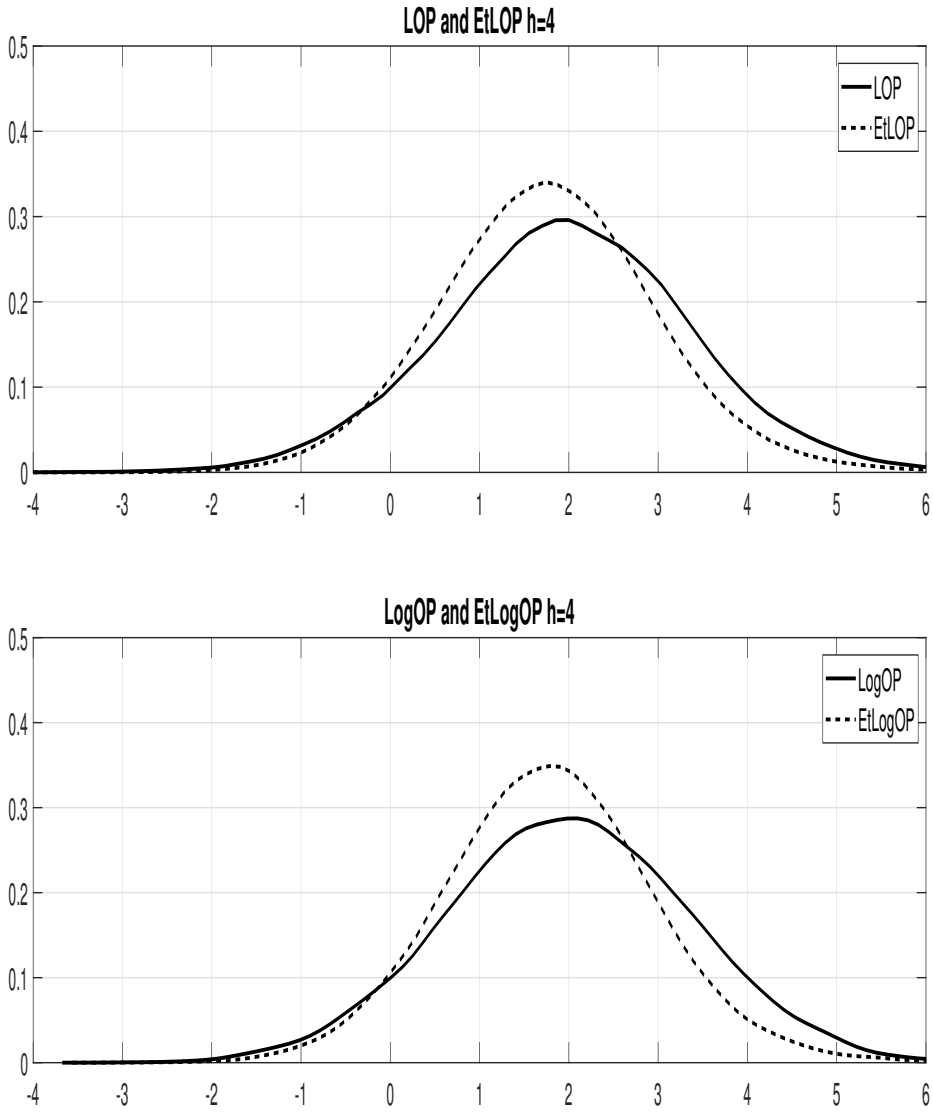
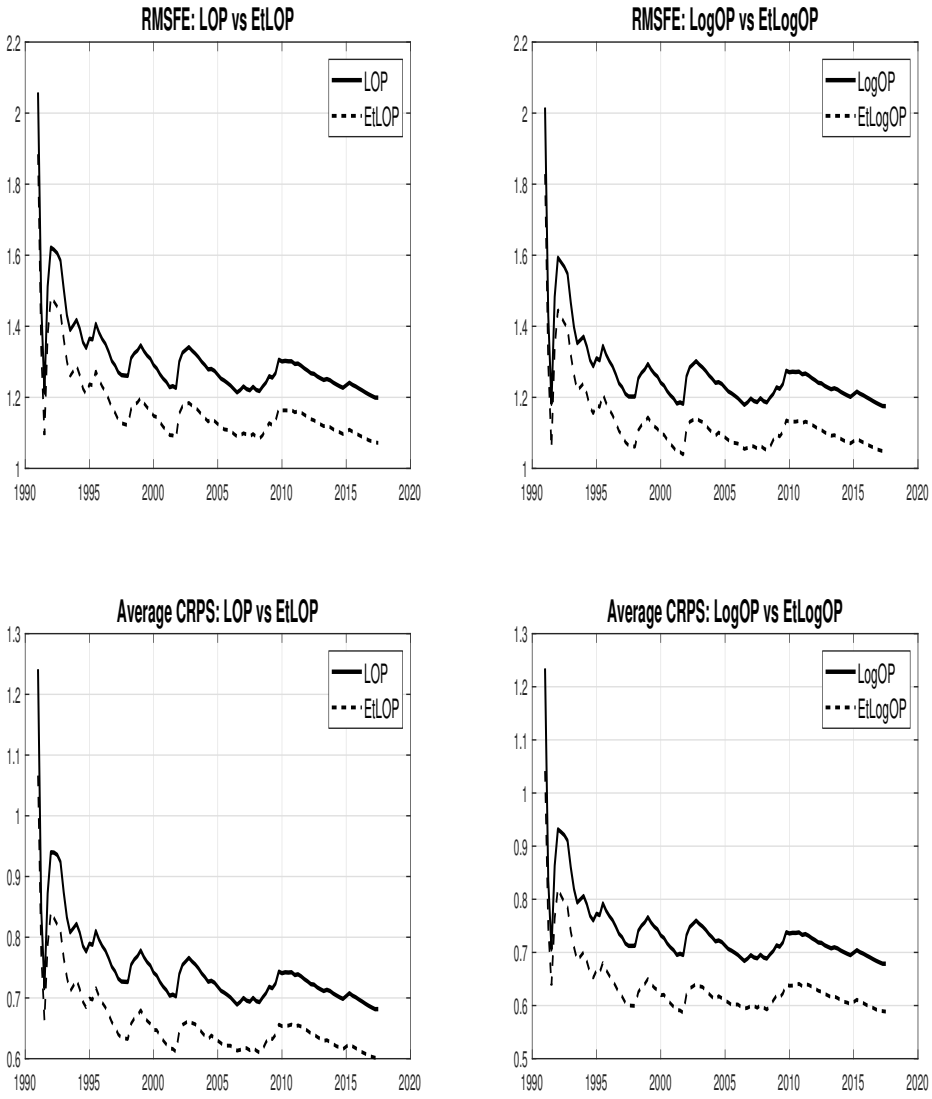


Figure 12: Recursive RMSFE and CRPS, $h = 4$



Appendix 3: Unobserved Components with Stochastic Volatility (UCSV)

Gauging the predictive accuracy of candidate methods relative to the UCSV approach has become standard within the inflation forecasting literature. With a focus on density forecasts, it is convenient to adopt the Bayesian approach described in Chan and Song (2018). The model specifies the following trend-cycle decomposition for inflation, π_t:

$$\pi_t = \pi_t^* + u_t^\pi, \qquad u_t^\pi \sim N(0, e^{h_t})$$

where π_t^* represents trend inflation and u_t^π is a transitory deviation from the trend, often referred to as the inflation gap. To define the UCSV model, we augment the trend-cycle decomposition with equations specifying AR(1) processes for the inflation trend, π_t^*, and the log volatilities of the transitory and trend components, h_t and g_t respectively:

$$\begin{aligned}
\pi_t^* &= \pi_{t-1}^* + u_t^{\pi^*}, \qquad u_t^{\pi^*} \sim N(0, e^{g_t}) \\
h_t &= h_{t-1} + u_t^h, \qquad u_t^h \sim N(0, \sigma_h^2) \\
g_t &= g_{t-1} + u_t^g, \qquad u_t^g \sim N(0, \sigma_g^2).
\end{aligned}$$

The model is estimated using Markov Chain Monte Carlo (MCMC) methods, implemented using the Gibbs sampler that sequentially draws from the full conditional distributions of the parameters and the latent states. The parameters here are σ_h^2 and σ_g^2 where the latent states are g, h and π^*. The priors are non-informative with estimated smoothing parameters σ_h^2 and σ_g^2; see Chan and Song (2018) Appendix A. We draw 50,000 iterates, with 5,000 burnin.