Federal Reserve Bank of Cleveland Working Paper Series

Censored Density Forecasts: Production and Evaluation

James Mitchell and Martin Weale

Working Paper No. 21-12R

August 2022

Suggested citation: Mitchell, James and Martin Weale. 2022. "Censored Density Forecasts: Production and Evaluation." Working Paper No. 21-12R. Federal Reserve Bank of Cleveland. https://doi.org/10.26509/frbc-wp-202112r.

Censored Density Forecasts: Production and Evaluation*

James Mitchell† and Martin Weale‡

August 15, 2022

Abstract

This paper develops methods for the production and evaluation of censored density forecasts. The focus is on censored density forecasts that quantify forecast risks in a middle region of the density covering a specified probability, and ignore the magnitude but not the frequency of outlying observations. We propose a fixed-point algorithm that fits a potentially skewed and fat-tailed density to the inner observations, acknowledging that the outlying observations may be drawn from a different but unknown distribution. We also introduce a new test for calibration of censored density forecasts. An application using historical forecast errors from the Federal Reserve Board and the Monetary Policy Committee (MPC) at the Bank of England suggests that the use of censored density functions to represent the pattern of forecast errors results in much greater parameter stability than do uncensored densities. We illustrate the utility of censored density forecasts when quantifying forecast risks after shocks such as the global financial crisis and the COVID-19 pandemic and find that these outperform the official forecasts produced by the MPC.

Keywords: Forecast uncertainty; Outliers; Fan charts; Skewed densities; Best critical region; Density forecasting; Censoring; Forecast evaluation
JEL Classification: C24; C46; C53; E58

*A previous version of this paper was titled: "Forecasting with Unknown Unknowns: Censoring and Fat Tails on the Bank of England's Monetary Policy Committee." We are grateful to four anonymous referees, and conference and seminar participants at the Bundesbank, ESCoE, Essex, the National Bank of Poland, Henley Business School (Reading), Strathclyde, WBS, LSE, SNDE (FRB Dallas), and FRB New York for helpful comments. Particular thanks for their comments to Jamie Bell, Todd Clark, Andrew Harvey, Ed Knotek, Malte Knüppel, Gary Koop, Paul Labonne, David Latto, John Maheu, Chuck Manski, Ivan Petrella, Barbara Rossi, and Simon van Norden. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Cleveland or the Federal Reserve System.

†Federal Reserve Bank of Cleveland; Economic Statistics Centre of Excellence (James.Mitchell@clev.frb.org)

‡King's College London; Economic Statistics Centre of Excellence (martin.weale@outlook.com)

1 Introduction

Many forecaster and central bank assessments of future uncertainties are informed, at least in part, by monitoring past forecast errors.[1] As Reifschneider and Tulip (2019) review, this is the general approach to gauging unconditional forecast uncertainty at the US Federal Reserve, the European Central Bank, the Bank of England, the Reserve Bank of Australia, the Bank of Canada, and the Swedish Riksbank.

Practice on communicating these forecast uncertainties varies. Famously, the Bank of England's Monetary Policy Committee (MPC) represents its forecast densities as fan charts, with shades of red (for inflation) and green (for GDP) representing regions with specified probabilities of outturns. Most of the analysis of these fan charts has drawn on the bank's specification of the underlying probability distribution, defined by a two-piece normal distribution, and devised evaluation tests for these density forecasts, making the assumption that the density function is described fully. But in fact - and this is commonly ignored - they describe only the inner 90 percent best critical region (BCR) of the forecast distribution. The BCR characterizes the interval of shortest length with a target (nominal) coverage rate of 90 percent. The MPC is effectively publishing what we call a censored density forecast. The censoring applies to both tails, although the degree of censoring can differ between the left and right tails.[2]

Censored density forecasts offer a way for the forecaster to quantify forecast risks in the middle of the density, ignoring the size but not the frequency of outlying forecast errors. This reflects, at a foundational level, a Knightian distinction between known and unknown probabilities: the distinction between *risk* and *uncertainty*. The censored region of the density acknowledges that there are (realized) unknown unknowns or events not expected to recur that should be censored before quantifying known unknowns. Censored densities also accord with the ideas formalized in Orlik and Veldkamp (2014) and Kozlowski, Veldkamp, and Venkateswaran (2020) that economic agents know more about the probabilities of everyday events than (black swan) events in the tails of a distribution,

[1] Central banks also use model-based approaches and their subjective judgment to gauge future economic uncertainties.

[2] In principle, censoring could apply to any region of the density. This paper focuses on censoring the tails, following practice at the MPC and our interpretation, discussed below, of what other central banks do.

given these are rarely observed.

Censored density forecasts, therefore, do not require the forecaster to quantify forecast uncertainties in the tails, beyond saying that there is, in sum, a 100α percent chance, where $\alpha \in (0, 1)$, of observing these more "extreme" events. Under density forecast asymmetry, we emphasize, this need not imply that $100(\alpha/2)$ percent of the probability mass falls in each tail. When the censoring bounds are shortest interval or best critical regions, the uncensored region need not lie between the $100(\alpha/2)$ percent and the $100(1 - \alpha/2)$ percent quantiles. Censored density forecasts require the forecaster to set these upper and lower censoring thresholds, providing them scope to indicate asymmetries in the balance of (unknowable) uncertainties. In contrast with value-at-risk (VaR) assessments, the censoring thresholds are unknown - and, as we shall explain, data determined - rather than fixed at a given quantile. Censored density forecasts provide a probabilistic impression of what will happen in the center of the distribution, whereas VaR assessments censor both this and the right-tail region of the density.

Many other central banks and official bodies also publish confidence intervals around their forecasts.[3] For example, the Congressional Budget Office (CBO) shows surrounding its forecast for the budget deficit a range that includes two-thirds of possible outcomes. Federal Reserve Board (FRB) staff forecasts, as presented historically in either the *Greenbook* or the *Tealbook* created before each meeting of the Federal Open Market Committee (FOMC), show a 70 percent confidence interval as does the Reserve Bank of Australia. The European Central Bank (ECB) has historically shown, around its GDP and inflation forecasts, a range of twice the mean absolute error, described by European Central Bank (2009) as consistent with a 57.5 percent confidence interval. These confidence intervals are generally calculated by making the assumption that the forecast errors are Gaussian.[4] In all of these examples, the forecast intervals are estimated based on past forecast performance. But the forecast intervals quantify only an inner proportion of the underlying density forecast. And, as we demonstrate, if the forecaster wishes to emphasize the distri-

[3]Appendix A in Tulip and Wallace (2012) helpfully summarizes the uncertainty measures used across various central banks.

[4]The FOMC in the *Summary of Economic Projections* emphasizes the approximate 70 percent interval. This explains why we interpret the FRB staff forecast intervals below as 70 percent intervals. Strictly, under Gaussianity with the length of the interval set as a one-standard-deviation error on each side of the point forecast, the intervals cover 68 percent. Results below are robust to analysis at 68 percent as opposed to 70 percent.

butional form of the inner confidence interval, it is possible to specify in parametric form the density function inside the confidence interval, without having any view on the distribution outside the confidence limits. Such an approach is particularly valuable when large outliers, such as those arising from the recent financial crisis and the COVID-19 pandemic, are present in the data.

In this paper we develop a new statistical approach to the problem of estimating and then evaluating censored densities for the errors associated with economic forecasts, or indeed with any set of observations. Our approach lies midway between a nonparametric approach and assuming that a parametric function applies to the whole of the distribution.[5] It requires neither subjective assessment of what constitutes an outlier nor parameterization of the outlier process. It involves fitting a parametric density to only the inner $100(1-\alpha)$ percent of observations, acknowledging that the outlying observations may be drawn from a different but unknown distribution. The technique is of obvious use in the aftermath of the current pandemic as an alternative to manual (judgment-based) adjustment.

Our approach thus offers an alternative to model-based methods such as those of Carriero et al. (2022), who, when modeling and forecasting macroeconomic data, downweight extreme observations, such as those observed during the COVID-19 pandemic, by allowing for both persistent and temporary heteroscedasticity. Transitory outliers are modeled either by t-distributed error processes or by parameterizing an outlier volatility state as in the so-called stochastic volatility outlier-adjusted (SVO) model of Stock and Watson (2016). Like our approach, the SVO model requires the forecaster to specify how frequently outliers occur. But unlike our approach, it requires making a choice about what density the outliers are drawn from. The use of censored density forecasts can also be contrasted with dropping outlier observations in estimation, as Schorfheide and Song (2020) propose when modeling in the aftermath of the pandemic.

The techniques needed for estimating censored distributions are well known. Observations in the censored region are given a likelihood computed from the probability of the observation in question being in the censored region. This method is easy to apply

[5]The Reserve Bank of Australia constructs its forecast error intervals using a nonparametric approach. It calculates its equal-tailed interval forecasts from the quantiles of the error distribution. As discussed, equal-tailed intervals need not equal BCR intervals for asymmetric distributions.

if the censoring is at known points. The probability of being in the censored region then depends on the mass of the uncensored part of the distribution lying between the known censor points. That is determined by the parameters describing the uncensored part of the distribution. But this method does not work if the censoring applies to a known proportion of the distribution rather than to observations outside known numerical limits. Accordingly, this paper develops a fixed-point algorithm to solve the problem of estimating the parameters of a distribution that applies to a known (central) proportion of a set of observations. We show that, for the GDP forecasts produced by the MPC and the FRB staff, this approach results in a distribution closer to Gaussianity than if all the observations were assumed to be distributed according to the parametric distribution.

The next section of the paper summarizes the forecast error data from the MPC and FRB used in the applications. Section 3 sets out the parametric family of skewed distributions that we consider, and, in the related Appendix A.2, their use in fitting FRB staff and MPC GDP growth forecast errors. We find that fat tails are present and, in the case of the MPC, pronounced skew. Section 4 sets out the implications for estimation when the density function is not fully described. We propose a fixed-point algorithm that fits the density by maximum likelihood, acknowledging endogenous censoring of the outlying 100α percent of observations. Section 5 explores the performance of this algorithm via Monte Carlo experiments suggesting that our algorithm performs well. Section 6 fits censored densities to the FRB and MPC GDP growth forecast errors, finding that the censored distributions are much more stable, in the face of large forecasting shocks like those associated with the pandemic, than are the uncensored distributions. Section 7 proposes and evaluates, via Monte Carlo methods, new calibration tests for censored density forecasts. These take account of both the proportion of outcomes in the censored region and the distribution of forecast errors in the uncensored region. Section 8 provides an out-of-sample evaluation of censored densities fitted to the FRB and MPC error data. In so doing we provide the first evaluation of the MPC's own density forecasts that correctly acknowledges the censoring. We find that censored distributions are more stable over time than are the uncensored distributions. We also show that, while the FRB produces satisfactory density forecasts based on a normal distribution whose parameters are estimated over a rolling 20-year window, censored density forecasts outperform the

subjective forecasts produced by the MPC. Section 9 concludes. The online Appendix contains supplementary details, results, and robustness checks.

2 Forecast Error Data

The forecast error data that we analyze in this paper are the MPC's and FRB staff's forecast errors for GDP growth. In the empirical analysis below we focus attention on the MPC forecasts at the two-year horizon and the FRB's forecasts at the one-year horizon.[6] The online Appendix (Section A.6) presents results for other forecasting horizons. As relevant, we summarize these below. Arguably, for an inflation-targeting central bank, the longer-horizon forecasts are of more interest. We consider forecasting errors relative to the modal forecast of the MPC and the point forecast issued by the FRB staff in the *Tealbook* (previously the *Greenbook*), widely believed to also be the mode, although, unlike the MPC, the FRB presents only symmetric density forecasts.

When calculating forecast errors for annualized quarterly GDP growth, since GDP data are revised, we need to decide which vintage of GDP data to use. Since the MPC explicitly sets out to forecast "mature" values of GDP, as noted in footnotes to its fan charts, for UK GDP growth we focus, for now, on using the latest available data vintage (from the quarterly national accounts published in February 2021) to define the outturn. This delivers GDP errors from 1999q4-2020q3, a time series of 84 observations. In Section 8, in out-of-sample analysis, we consider the use of second-release GDP data to measure the outturn. For the FRB, the staff forecasts are released with a five-year lag. We consider their projections for quarter-on-quarter growth in real GDP (annualized percentage points) from 1974q2-2014q4, a time series of 163 observations. Again the outturn needs to be defined, and following Reifschneider and Tulip (2019), we define it using data published soon after the release of the forecast. Specifically, we define the outturn for US GDP as the second-release GDP estimate. For the US, unlike the UK, the use of mature rather than second-release data to define the outturn has little effect on either the shape of the forecast error histograms or the fitted densities seen below.[7]

[6] The FRB has not always forecast GDP growth two years ahead. Hence, we focus here on the longer and unbroken sample of one-year-ahead forecasts.

[7] Compare online Figures A2 and A8.

3 The Distribution of Forecast Errors: A Parametric Framework

The aftermath of the global financial crisis saw increased attention paid to forecast errors and their distributional form; for example, see Alessi et al. (2014). Haldane (2012) noted how theory and evidence suggest that macroeconomic data exhibit fat tails as well as skewness, and Adrian, Boyarchenko, and Giannone (2019) emphasize non-Gaussian features when measuring the "vulnerability" of GDP growth to downside risks. Using long-run historical data, Jordà, Schularick, and Taylor (2020) conclude that growth is pervasively fat tailed and non-Gaussian.

It is clear that the 2020 pandemic has again raised questions about how forecast errors should be treated. A related literature considers how models used to produce density forecasts should be designed, given the extreme data realizations and consequent forecasting errors observed during the pandemic; for example, see Schorfheide and Song (2020), Carriero et al. (2022), Lenza and Primiceri (2022), and Huber et al. (forthcoming).

We focus, instead, on the construction of density forecasts - and in due course censored density forecasts - from historical forecast error data. We limit our consideration to forecasts based on density functions whose uncensored regions are unimodal and continuous, with the density strictly increasing to the left of the mode and strictly decreasing to the right. We also assume that the derivatives of the density function with respect to its parameters all exist and that the parameters do not take boundary values. This is consistent with the assumptions of bodies such as the MPC and the FRB and ensures that the best critical region, i.e., the smallest range containing the required probability mass, for any probabilistic forecast, is both continuous and unique (Turkkan and Pham-Gia, 1997). These restrictions do, however, allow us to consider density functions that allow for fat tails and skewness.[8] The density functions we adopt nest the normal and the two-piece normal densities used by the Federal Reserve Board and the Bank of England,

[8]While we maintain an assumption of unimodality, as we shall see below additional modes are permitted in the censored region. Hence, the researcher always has the scope to censor more of his or her density to ensure that the uncensored region is unimodal. Therefore, we encourage any user to look at his/her data prior to fitting a censored density forecast. If the data are truly multimodal, with distinct and significant modes, then the rationale for fitting a censored density forecast is diminished. Our algorithm is designed for situations - as explained not uncommon in macroeconomics - when the data are unimodal, albeit perhaps skewed and/or fat-tailed, within a central region.

respectively. Specifically, we consider the general family of skew distributions defined in Arellano-Valle, Gómez, and Quintana (2005). But the analysis of censored densities that follows could be performed for any parametric density function.[9]

Like the two-piece normal, the skew distribution of Arellano-Valle, Gómez, and Quintana (2005) involves joining two distributions, with different scale (and perhaps shape) parameters. We return to the issue, central to this paper, that the forecaster may have no view on the outer 100α percent of the distribution in Section 4.

A leading specific density within this family that we focus on is the two-piece t distribution described by Fernandez and Steel (1998). This depends on, in addition to the location, scale, and skew parameters, the number of degrees of freedom of the t distribution. For robustness, in online Appendix A.1 we explore more general and alternative skewed specifications for the MPC forecast errors, given that, as we shall see, these data are less symmetric than the FRB error data. In general, we find that (in-sample) the two-piece t fits the MPC data competitively relative to these alternatives. We therefore confine our attention to it (and its limiting case, the two-piece normal distribution) here.

The density function of the two-piece t, $2Pt(\nu,\mu,\sigma,\gamma)$, is given as:

$$f(y_t) = \frac{2}{\sigma\left(\gamma + 1/\gamma\right)} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{1/2}} \left[1 + \frac{(y_t-\mu)^2}{\gamma^2\nu\sigma^2}\right]^{-(\nu+1)/2} \quad \text{if } y_t < \mu \tag{1}$$

$$f(y_t) = \frac{2}{\sigma\left(\gamma + 1/\gamma\right)} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{1/2}} \left[1 + \frac{\gamma^2(y_t-\mu)^2}{\nu\sigma^2}\right]^{-(\nu+1)/2} \quad \text{if } y_t \geq \mu.$$

where $\gamma \in (0,\infty)$ is the scalar skew parameter, $\nu > 0$ is the degrees of freedom of the standard Student t distribution with location, μ, and scale, $\sigma > 0$, and $\Gamma(.)$ is the gamma function.

The mode of the distribution is μ but this is the same as the mean only if $\gamma = 1$. The probability mass to the left of the mode is $\gamma^2/(\gamma^2+1)$, while that to the right of the mode is $1/(\gamma^2+1)$. So with $\gamma < 1$ the distribution is skewed to the right and with $\gamma > 1$ it is skewed to the left. A large number of degrees of freedom, ν, implies, of course, that the

[9]We leave for future research the issue of how censored density forecasts should be produced from workhorse macroeconomic forecasting models, such as vector autoregressive (VAR) models. For linear (parametric) VAR models this will require assumptions about the distributional form of the errors to the model to apply only within censoring bounds. Nonparametric VAR models, such as those developed by Huber et al. (forthcoming), may offer more flexibility.

distribution is very close to normal, while a small number of degrees of freedom indicates that extreme values are appreciably more common than would be implied by a normal distribution with the same scale parameter.[10]

Given a scoring rule or loss function the parameters of this distribution can be estimated. Following Gneiting and Raftery (2007), optimum score estimators or M-estimators involve maximizing the value of the (proper) scoring rule over the sample. We focus on the logarithmic scoring rule corresponding to maximum likelihood (ML) estimation.[11]

The log-likelihood function of a sequence of observations y_t, $t = 1, ..., T$, where $\mathbf{I}(y) = 1$ if $y \geq 0$ and $\mathbf{I}(y) = 0$ if $y < 0$, is given as:

$$\log L = T \ln \left(\frac{2}{\sigma\left(\gamma + 1/\gamma\right)} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\left(\pi\nu\right)^{1/2}} \right) + \sum_{t=1}^{T} \mathbf{I}\left(y_t - \mu\right) \ln \left[1 + \frac{\gamma^2 (y_t - \mu)^2}{\nu\sigma^2} \right]^{-(\nu+1)/2} + \sum_{t=1}^{T} \mathbf{I}\left(\mu - y_t\right) \ln \left[1 + \frac{(y_t - \mu)^2}{\gamma^2 \nu \sigma^2} \right]^{-(\nu+1)/2}. \quad (2)$$

For any given sample, the four parameters, μ, σ, γ, and ν can be estimated by ML. Our sample suffers from the drawback that the forecast errors relate to overlapping periods. Nevertheless, parameter estimation by ML delivers consistent estimates (White, 1980).

In Section A.2 of the Appendix we show the results of fitting the skewed t distribution to forecast errors for GDP and inflation. We find that, despite their greater flexibility, the t distributions still appear to have trouble accommodating the extremes of the distribution. If, instead of being fitted to the whole distribution, they were fitted only to the central part, one might expect to see less skew and perhaps a t distribution closer to normality. A normality assumption, for the central region of the density, might be appropriate after all. We explore this issue next by fitting censored two-piece t and normal distributions to these forecast errors.

[10] In this specification the scale parameter to the left of the mode is $\sigma\gamma$, while to the right of the mode it is σ/γ. In the specification used by the MPC it is $\left(\frac{\sigma^2}{1-\phi}\right)^{1/2}$ to the left of the mode and $\left(\frac{\sigma^2}{1+\phi}\right)^{1/2}$ to the right of the mode. So it is easy to express γ in terms of ϕ and *vice versa*.

[11] The logarithmic score is known to be more sensitive to outliers than alternatives such as the continuous ranked probability score (CRPS). Future work might consider estimators that minimize CRPS loss along the lines of Gebetsberger et al. (2018).

4 Fitting Censored Distributions

There are different types of forecast intervals, which implies that one can construct different types of censored density forecasts. We focus our discussion on the construction of BCR or shortest-interval censored density forecasts. They represent the shortest range of possible outcomes that have the required probability.[12] In a decision theory framework, Wallis (1999) and Askanazi et al. (2018) show that the shortest interval is the *best* prediction interval when the loss function takes an all-or-nothing form. This is such that the loss (or cost) of an outturn falling outside the BCR in question is the same irrespective of how far away from the BCR the outturn falls.

But we do briefly discuss the equal-tailed censored density that lies between the $100(\alpha/2)$ percent and the $100(1-\alpha/2)$ percent quantiles. Equal-tailed intervals differ from BCR forecasts under asymmetry of the underlying density. Askanazi et al. (2018), Brehmer and Gneiting (2021), and Taylor (2021) provide recent analysis comparing alternative ways to produce and evaluate such interval forecasts. We should stress our points of departure from these authors. Unlike them, we quantify the density function within the interval. And in evaluation, we assume that the censoring quantiles are known; hence, the elicitability problems emphasized by Askanazi et al. (2018) and Brehmer and Gneiting (2021) for BCR intervals are not present.

The principles of fitting BCR censored distributions when the censor points are given exogenously are well understood. Typically it is clear whether observations are censored or not, but not where they lie in the censored region. In that situation, Diks, Panchenko, and van Dijk (2011) have shown that in computing the likelihood function, the censored observations are given a likelihood equal to the chance of being in the censored region, conditional, of course, on the parameters of the distribution. This yields ML estimates of

[12] To a Bayesian, the $100(1-\alpha)$ percent best critical region/interval for y is the highest posterior density (HPD) interval:

$$R_\alpha = \{y : f(y) \geq \pi_\alpha\}, \text{ where}$$
$$\pi_\alpha \text{ is the largest value for which } P(y \in R_\alpha) \geq 1-\alpha$$

An HPD interval has two main properties: (1) the density for every point inside the interval is greater than that for every point outside the interval and (2) for a given probability the interval is of shortest length; see Hyndman (1996) for methods to estimate HPD intervals.

the parameters, with standard properties.[13]

Although we use the censored likelihood function of Diks, Panchenko, and van Dijk (2011), the situation we face is different in two respects, given our interest in forecast production rather than just evaluation. First, while we require observations outside the $100(1-\alpha)$ percent BCR (or shortest interval), we do not wish their position to have any influence on the estimated parameters of the distribution. This can be achieved if they are treated as though they are censored with a likelihood defined by the probability, 100α percent, of being in the censored region. Thus, conditional on known censor points, this difference is not material for estimation.

The second difference, however, is material. In the situation we face, the censor points are defined by the bounds of the $100(1-\alpha)$ percent BCR and thus by the parameter estimates. The "regularity conditions" needed to prove, in particular, the asymptotic normality of ML estimators are violated because the support of the density depends on its parameters, as in (3); for example, see Woodroofe (1972) and Smith (1985). We now show that, in our case, use of the standard ML estimator produces degenerate results and we develop an alternative fixed-point algorithm to produce our parameter estimates.

4.1 Motivating a Fixed-Point Algorithm

If the lower cut point, beyond which data are censored, is y_L and the upper cut point, above which data are censored, is y_U, then following Diks, Panchenko, and van Dijk (2011) the conventional way of setting out the censored log likelihood, which we refer to as L_A^C, is:

$$\log L_A^C = \left\{ \begin{array}{ll} \log(F(y_L)) & \text{if } (y < y_L) \\ \log L & \text{if } (y_L \leq y \leq y_U) \\ \log(1 - F(y_U)) & \text{if } (y > y_U) \end{array} \right\}, \tag{3}$$

[13] Holzmann and Klar (2017) further analyze and confirm the good properties of the censored likelihood of Diks, Panchenko, and van Dijk (2011) in forecast evaluation (as opposed to forecast production). They also propose alternative weighted scoring rules that focus forecast evaluation on regions of specific interest. These could be of interest if the forecaster did wish to emphasize specific regions within the uncensored region of the censored density forecast.

where $F(y)$ defines the CDF of the density function, $F(y) = \int_{-\infty}^{y} f(y)dy$, and the BCR, set to define a 100α percent censored region, satisfies:

$$F(y_U, \beta) - F(y_L, \beta) = 1 - \alpha. \tag{4}$$

$$f(y_U, \beta) - f(y_L, \beta) = 0 \tag{5}$$

Condition (4) ensures that the uncensored region is of the required size. Given that we consider only unimodal distributions whose mode lies in the uncensored region, equation (5) results from minimization of $y_U - y_L$ subject to the constraint of (4). This ensures that the uncensored region is the smallest possible region that contains the requisite probability mass.[14]

These conditions, together with the derivatives of the log-likelihood with respect to β, ensure that we have as many equations as there are unknowns: the elements of β and the two cut points. This likelihood function, however, still assumes that the forecaster has a view on whether points are likely to be in the upper or the lower tail of the distribution, notwithstanding that the density function within those tails is not specified.

An alternative likelihood function, L_B^C, which is completely agnostic as to whether observations are going to be above the upper cut point or below the lower cut point, can be defined, with conditions (4) and (5) again imposed, as:

$$\log L_B^C = \left\{ \begin{array}{ll} \log(1 - (F(y_U) - F(y_L)) & \text{if } (y < y_L) \\ \log L & \text{if } (y_L \leq y \leq y_U) \\ \log(1 - (F(y_U) - F(y_L)) & \text{if } (y > y_U) \end{array} \right\}. \tag{7}$$

We subsequently show the advantages in estimation, especially in smaller samples, of the greater structure provided by L_A^C. But it should be noted that neither condition (4) nor (5) require that a proportion α of the observations lie in the censored region.

As a result, in either case, estimation of $\beta = [\mu, \sigma, \gamma, \nu]$, subject to (4) and (5), is difficult, indeed potentially degenerate. This is because the censor points are treated as endogenous.

[14] We note that substitution of (5) for:

$$F(y_L, \beta) = \alpha/2 \tag{6}$$

delivers an equal-tailed, as opposed to BCR or shortest-interval, censored density forecast.

Intuitively, for fixed (finite) T, we explain the degeneracy of standard ML results when estimating β, y_L, and y_U as follows. We illustrate for L_A^C, although the same point is pertinent for L_B^C. Consider:

$$\max_{\beta, y_L, y_U} \sum \log L_A^C(y_t, \beta) + \lambda_1 \left(F(y_U, \beta) - F(y_L, \beta) - (1-\alpha)\right) \\ + \lambda_2 \left(f(y_U, \beta) - f(y_L, \beta)\right). \tag{8}$$

Suppose that we have a value of σ sufficiently small such that only one observation from a sample, say, y_A, is in the uncensored region and that this is the value given to the mode of the distribution, μ: $\mu = y_A$. All other observations are then in the censored region - with, say, T_1 observations below y_L and T_2 observations above y_U. Then, when the constraints are met:

$$\log L_A^C = T_1 \log F(y_L, \beta) + \log f(y_A, \beta) + T_2 \log(1 - F(y_U, \beta)). \tag{9}$$

But as σ shrinks, for fixed T, $\log f(y_A, \beta)$ will increase without limit:

$$\log L_A^C \to \infty \text{ as } \sigma \to 0. \tag{10}$$

In the absence of censoring this would be offset by the likelihood associated with the other observations falling. But with the censored likelihood, for fixed T, that is not the case. In other words, the censor points y_L and y_U change as σ shrinks, but the probability of being in the censored tails, and thus $F(y_U, \beta)$ and $F(y_L, \beta)$, will not change. For fixed T the overall log likelihood, $\log L_A^C$, is therefore unbounded as σ shrinks to zero; there is no interior solution.[15]

Accordingly, we suggest the following fixed-point algorithm in finite samples. It is motivated by the observation that, in large samples, estimates (for β) produced by maximizing $\log L^C$, with *fixed* censor points, are independent of the censor points, provided all the uncensored observations are genuinely drawn from the specified distribution.

[15] A similar unboundedness arises from the use of standard ML algorithms when estimating mixture densities with heterogeneous variances: the likelihood goes to infinity as the variance of one of the component densities goes to zero. The mixture literature has adopted alternative solutions to this problem including estimators on constrained parameter spaces and penalized estimators (for example, see Chen, Tan, and Zhang (2008)), that can also be interpreted and developed within a Bayesian framework (for example, see Hamilton (1991)). Variants of these approaches may also prove effective in our context and are left as a topic for future research. As explained subsequently, our solution to unboundedness in this paper is an algorithm motivated by the observation, unique to censored density estimation, that ML is valid for exogenous censoring intervals.

The proposed fixed-point algorithm contains the following two steps:

$$\text{Step 1: } \beta_{r+1} = \arg\max_{\beta} \sum \log L_j^C(y_t, \beta, y_{L,r}, y_{U,r}) \tag{11}$$

$$\text{Step 2: compute BCR of } f(y_t \mid \beta_{r+1}) \Rightarrow y_{L,r+1}, y_{U,r+1} \tag{12}$$

$j \in \{A, B\}$, where we search over values of $y_{L,r}$ and $y_{U,r}$ $(r = 1, ..., R^*)$ to minimize $P_{r+1} = (y_{L,r+1} - y_{L,r})^2 + (y_{U,r+1} - y_{U,r})^2$. If P_{r+1} converges to zero as R^* increases, this provides a solution at which the ML estimates of the parameters of the censored distribution deliver censor points that, when used in estimation, deliver the same parameter estimates. The derivative conditions and conditions (4) and (5) are met. We note that taking the censor points as fixed, ML estimation produces consistent parameter estimates. The censor points generated by the distribution are therefore, as functions of consistent parameter estimates, also consistent, conditional on the earlier set of censor points being correct. Repeated use and testing of the algorithm suggests that, given reasonable initial values,[16] it converges to a local maximum at which the proportion of observations in the censored region is α. This contrasts with the degenerate global maximum at which only one observation is in the uncensored region. Iterating over the fixed points reaches the desired solution because the algorithm does not explore the full range of possible parameter values in the way that a standard ML algorithm does.

The contribution of each observation to the log likelihood depends on whether it is in the censored region or not. The log likelihood will not be continuous in the parameters because, for some parameter sets, observations may be uncensored, while for others they will be censored. In large samples this effect is likely to be small; the contribution of each observation to the total log likelihood is low. But in small samples the discontinuities will be relatively greater and it may not be possible to find a solution for which the quadratic term converges to zero. If the minimum $P_r = (y_{L,r+1} - y_{L,r})^2 + (y_{U,r+1} - y_{U,r})^2 > 0$, only an approximation will have been found. It has to be a matter of judgment as to how good or bad that approximation is.[17]

In practice, in our experiments we found that, especially in moderate samples, maximization of L or L_j^C, $j \in \{A, B\}$, (for fixed censor points) could prove problematic: the

[16] We obtain β_1 by fitting the uncensored skewed t distribution to the full set of observations.

[17] We found convergence typically tends to occur for $R^* < 20$, which takes fewer than 30 seconds in Matlab on a standard desktop computer.

ML estimates of γ can diverge. Similar findings are reported by Sartori (2006) and Azzalini and Arellano-Valle (2013) for their skew normal and t densities (considered in more detail in Appendix A.1). This is because the likelihood can be monotone and the Fisher information matrix singular at the discontinuity point when skewness disappears, $\gamma = 1$. Accordingly, when sample sizes are small, in the spirit of Sartori (2006) and Azzalini and Arellano-Valle (2013), to avoid boundary estimates we suggest in Step 1 maximization of a penalized log-likelihood function, $PL_j^C(y_t, \beta)$, rather than L_j^C, where for $j \in \{A, B\}$

$$PL_j^C(y_t, \beta) = \sum \log L_j^C(y_t, \beta) - \frac{1}{2} P_\lambda(|(\gamma - 1)|) \tag{13}$$

and $P_\lambda(|(\gamma - 1)|)$ is a nonnegative penalty function. We use the Lasso penalty, $P_\lambda(|(\gamma - 1)|) = \lambda\,|(\gamma - 1)|$, where λ is a tuning parameter. When $\lambda = 0$ estimation reduces to $L_j^C(y_t, \beta)$; the higher the value of λ the more deviations from symmetry are penalized. In the Monte Carlo experiments that follow, we select λ by optimizing the in-sample censored fit, $\sum \log L_j^C(y_t, \beta_{R^*})$. We also experimented with the use of this penalized estimator in the empirical application of Section 8, but found that no penalty was required for satisfactory estimation and convergence of the fixed-point algorithm, that is, the estimates of $\lambda = 0$.[18]

The conditions imposed on the density function in Section 3, unimodality and a density that is strictly increasing or decreasing except at the mode, are sufficient to ensure that there is a unique best critical region associated with any set of parameter values for the density function. Equation (4) and condition (5) do not rule out the possibility of multiple solutions and, given that the density function is discontinuous at the cut points, we cannot rely on the results of Cox (2020). In the presence of multiple nondegenerate solutions, it would be natural to choose the parameters consistent with the global nondegenerate maximum. But we have found no evidence of multiple solutions.

5 Monte Carlo Evidence

We carry out Monte Carlo experiments to assess the performance of the proposed fixed-point algorithm, (11)-(12), in samples of different sizes as the degree of skew varies. We

[18] Note the connection between the use of $PL_j^C(y_t, \beta)$ and Bayesian *a posteriori* estimates with a Laplace prior on $(\gamma - 1)$.

make comparison with the penalized estimator, (13). We focus on censoring at 100α=10 percent, in keeping with MPC practice.

Here we evaluate the fixed-point algorithm in the central case when not all of the underlying data are drawn from the same (skewed) distribution: in particular when what will be identified as the censored observations come from a different distribution.

The online Appendix contains two additional sets of experiments that we summarize here. The first experiment finds that the fixed-point algorithm does well, relative to the uncensored estimator, at recovering the parameters of skewed densities for larger samples. But in small samples and when the underlying density is highly skewed, we find benefits to using the penalized censored estimator, with L_A^C also outperforming L_B^C. The second experiment assesses whether any parameter estimates produced when fitting the censored two-piece t to the time series of forecast errors could have been, in reality, generated by an underlying symmetric normal distribution. The experiments allow for serial correlation in forecast errors as when multi-step-ahead forecasting. As well as again evidencing the benefits of L_A^C versus L_B^C, we find that when the data are censored so that the distribution is fitted only to the central 90 percent of observations, for samples of error data of the length seen for the UK where the forecast error data are skewed, the estimated value of the number of degrees of freedom has to be 2.7 (2.2) or lower under L_A^C (L_B^C) before one can reject, at a 90 percent significance level, the hypothesis that the underlying distribution is normal. We refer to this result below.

5.1 Performance for Mixed Distributions

We explore the performance of the censored estimators under both L_A^C and L_B^C, when not all of the underlying data are drawn from the same (skewed) distribution: in particular, we allow (what will be) the censored observations to be drawn from a different distribution. T observations are first drawn from a two-piece t distribution, where $(\nu, \mu, \sigma, \gamma) = (5, 0, 1, 1.5)$ and $(5, 0, 1, 2.5)$. But then each of the 10 percent of these T observations that falls outside the 90 percent BCR defined by y_L and y_U, as estimated for each replication, is dropped and replaced, depending on whether it falls below y_L or above y_U, with a random draw from a uniform density between -10 and y_L or y_U and

+10, respectively.[19] Comparison is made with the uncensored ML estimator, L. We also report results for $PL_j^C(y_t, \beta)$, $j \in \{A, B\}$. We do not report (for space reasons) results for $PL_j^C(y_t, \beta)$ when $\gamma = 1.5$, since, as will be seen, the utility of the penalized estimators, relative to the unpenalized ones, is found to be greater in populations with high skew.

The mean, median, and standard deviation (across the $1,000$ replications) of the estimates of the four parameters are shown in Table 1.[20] We also report the proportion (averaged across the R replications) of the T observations that, for the censored estimators, are classified as falling in the censored region. Table 1 reports results for $T = 40, 100, 500,$ and $1,000$, noting the Bank of England's use of just 40 error observations to estimate its error densities. Let us consider the larger sample results first. When $T = 1,000$, we find that the censored estimators, again especially L_A^C, do a good job of estimating the true parameter values, despite the censoring. They also correctly place 10 percent of observations in the censored region. But, as expected, the uncensored estimator - which assumes that all T observations come from a single density - is not able to return estimates that are as accurate. It tends to overestimate $1/\nu$ (that is, underestimate ν), in an attempt to capture the 10 percent of tail observations drawn from the uniform densities.

As T decreases and γ increases, we again observe a higher chance that the censored estimates for γ diverge for some replications: as the mean estimates for γ again become too large, with the standard deviation estimates for γ elevated. Table 1 shows that in smaller samples this afflicts L_B^C more than L_A^C. The median estimates for L_A^C are closer to the true parameter values than the mean ones, especially so for smaller T. For $T = 40$ and $\gamma = 2.5$, focusing on the median estimates for γ, L_A^C is considerably more accurate than L_B^C, with L_B^C again tending, for an increasing number of replications, to overestimate γ (and underestimate σ). Use of the penalized estimator mitigates this small-sample concern further. The median estimates for the penalized estimator, under L_A^C, are within 10 percent of the true parameter estimates when $T = 40$. But it does not eliminate the risk that the skewness estimates will diverge, as the mean estimates from PL_A^C still diverge, suggesting that in any specific application with small samples care should be exercised, and parameter estimates closely inspected, if boundary values are to be avoided. The uncensored estimator continues to overestimate $1/\nu$ in small samples.

[19] Stock and Watson (2016), for example, also use the uniform density to model outliers.

[20] To mitigate computational issues when $\nu \to \infty$, we found it helpful to work with $1/\nu$ rather than ν.

Table 1: Monte Carlo results assessing the fixed-point estimator: Mean, median, and standard deviation (across replications) of the parameter estimates from the censored estimators L_A^C, L_B^C, PL_A^C, and PL_B^C and the uncensored ML estimator, L; and proportion (Prop) of observations placed in the censored region

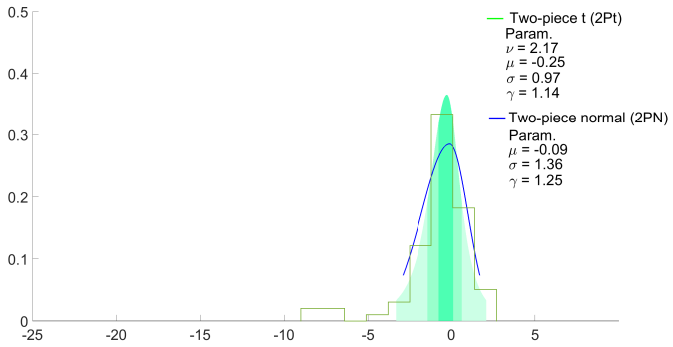
		T=40					**T=100**					**T=500**					**T=1,000**				
		$1/\nu$	μ	σ	γ		$1/\nu$	μ	σ	γ		$1/\nu$	μ	σ	γ		$1/\nu$	μ	σ	γ	
true		**0.20**	**0.00**	**1.00**	**1.50**	**Prop**	**0.20**	**0.00**	**1.00**	**1.50**	**Prop**	**0.20**	**0.00**	**1.00**	**1.50**	**Prop**	**0.20**	**0.00**	**1.00**	**1.50**	**Prop**
L_A^C	**mean**	0.37	-0.02	0.83	369	0.10	0.33	-0.01	0.91	2.01	0.10	0.26	0.00	0.97	1.51	0.10	0.24	0.00	0.98	1.50	0.10
	med	0.32	0.01	0.84	1.52	0.10	0.31	-0.02	0.90	1.49	0.10	0.24	0.00	0.97	1.51	0.10	0.23	0.00	0.97	1.50	0.10
	sd	0.35	0.49	0.27	1E+4	0.02	0.26	0.30	0.17	10.0	0.01	0.14	0.12	0.08	0.12	0.00	0.11	0.08	0.06	0.09	0.00
L_B^C	**mean**	0.27	0.01	0.60	1E+5	0.12	0.24	0.09	0.83	43.7	0.11	0.21	0.07	0.96	1.66	0.11	0.21	0.05	0.97	1.61	0.10
	med	0.00	0.13	0.64	1.85	0.10	0.15	0.07	0.87	1.64	0.11	0.21	0.06	0.95	1.59	0.10	0.21	0.05	0.96	1.58	0.10
	sd	0.37	0.78	0.95	4E+6	0.10	0.28	0.53	0.30	655	0.06	0.14	0.19	0.11	0.36	0.01	0.10	0.12	0.08	0.19	0.01
L	**mean**	0.49	-0.05	0.78	3.53		0.51	-0.06	0.83	1.52		0.52	-0.06	0.85	1.45		0.53	-0.07	0.85	1.44	
	med	0.50	-0.02	0.78	1.50		0.52	-0.07	0.83	1.44		0.52	-0.06	0.85	1.44		0.53	-0.07	0.85	1.44	
	sd	0.21	0.49	0.22	12.09		0.12	0.30	0.12	0.43		0.05	0.11	0.05	0.12		0.04	0.08	0.04	0.08	
true		**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**
L_A^C	**mean**	0.30	-0.06	0.72	2E+5	0.10	0.26	-0.01	0.88	709	0.10	0.23	0.00	0.98	2.55	0.10	0.22	-0.01	0.99	2.51	0.10
	med	0.24	0.01	0.80	2.65	0.10	0.24	0.00	0.92	2.53	0.10	0.22	0.00	0.97	2.50	0.10	0.22	0.00	0.98	2.50	0.10
	sd	0.29	0.56	0.51	4E+6	0.05	0.21	0.35	0.32	2E+4	0.02	0.10	0.13	0.11	0.47	0.00	0.07	0.08	0.07	0.20	0.00
L_B^C	**mean**	1E+3	1E+3	0.54	5E+6	0.14	0.18	0.14	0.58	692	0.11	0.19	0.08	0.89	9.84	0.10	0.20	0.05	0.94	3.62	0.10
	med	0.00	0.19	0.03	51.57	0.10	0.00	0.19	0.67	3.98	0.10	0.20	0.06	0.92	2.74	0.10	0.21	0.04	0.94	2.66	0.10
	sd	4E+4	4E+4	4.32	2E+8	0.18	0.23	0.42	0.48	5E+3	0.07	0.12	0.20	0.23	58.6	0.01	0.08	0.13	0.12	26.60	0.00
L	**mean**	0.30	-0.13	0.77	18.27		0.33	-0.15	0.92	5.50		0.36	-0.17	0.99	2.22		0.37	-0.17	1.00	2.19	
	med	0.27	0.00	0.85	2.67		0.35	-0.13	0.95	2.25		0.36	-0.16	0.99	2.20		0.37	-0.17	1.00	2.17	
	sd	0.21	0.59	0.48	30.89		0.14	0.40	0.29	13.7		0.05	0.15	0.09	0.28		0.03	0.10	0.06	0.18	
true		**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**
PL_A^C	**mean**	0.25	-0.11	0.81	7E+4	0.11	0.24	-0.06	0.94	2E+3	0.11	0.23	-0.05	1.01	2.43	0.10	0.22	-0.04	1.01	2.43	0.10
	med	0.18	-0.01	0.89	2.55	0.10	0.23	-0.05	0.95	2.43	0.11	0.23	-0.05	1.01	2.41	0.10	0.22	-0.04	1.01	2.41	0.10
	sd	0.27	0.55	0.45	2E+6	0.03	0.20	0.33	0.28	5E+4	0.02	0.11	0.13	0.10	0.28	0.00	0.08	0.09	0.08	0.20	0.00
PL_B^C	**mean**	0.15	-0.14	0.77	4E+4	0.11	0.15	-0.09	0.94	158	0.11	0.15	-0.03	1.02	3.41	0.11	0.17	-0.04	1.02	2.56	0.11
	med	0.00	0.08	0.78	3.18	0.10	0.00	-0.01	0.97	2.58	0.11	0.16	-0.01	1.00	2.52	0.11	0.19	-0.02	1.00	2.50	0.10
	sd	0.25	0.74	0.48	9E+5	0.04	0.21	0.45	0.41	3E+3	0.03	0.14	0.21	12.7	12.7	0.01	0.11	0.16	0.16	0.87	0.01

6 Censored Densities Fitted to the MPC and FRB Forecast Errors

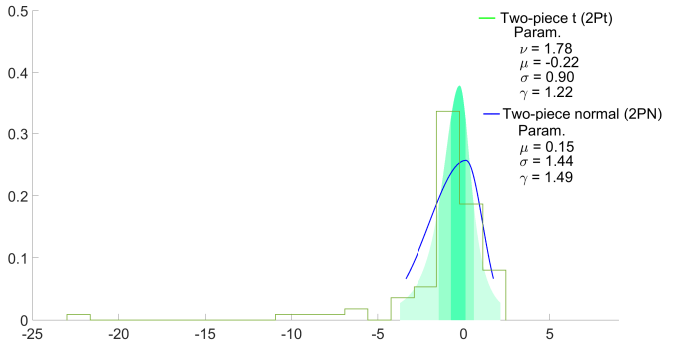
We now fit the censored density functions to the MPC and FRB staff forecast error data shown as histograms in Figures 1 and 2. We use L_A^C because, consistent with the Monte Carlo evidence and attempts to fit L_B^C to the forecast errors (reported in Section A.8.1 of the online Appendix), L_A^C better avoids boundary solutions for γ when fitting the two-piece normal distribution. In all cases $P_r = 0$ (for large r), confirming satisfactory estimation. The online Appendix (see Tables A5 and A6) provides supplementary results for other forecast horizons and when censoring the MPC's density forecasts at 30 percent rather than 10 percent, as in practice, and *vice versa* for the FRB. We refer to relevant aspects of these supplementary results in the discussion below. In summary, as expected, these tables reveal that the variances of the uncensored Gaussian forecast error densities increase with the forecast horizon. But this is often not the case for the skewed and fat-tailed densities (when they differ from the Gaussian density, which, as seen, they do for the UK but not the US sample). The skew of the uncensored densities tends to increase with the forecast horizon in the UK but decrease in the US.

Censored forecast error densities at the 10 percent level are shown for the UK in Figure 1 and for the US, censoring at 30 percent, in Figure 2. In each case, pre-pandemic densities are shown in panel (a) and densities including the pandemic in panel (b). The US densities in panel (b) involve appending to the actual forecast error data a single artificial observation of -20 percent to simulate the likely effects of the forecast errors made during COVID-19. Table 2 reports the parameter estimates of the censored densities, alongside the estimates when an uncensored density is fitted.

Looking at the UK first (pre-pandemic) and comparing Figure 1 with the uncensored distributions of Figure A1, we see that not allowing the outlying 10 percent of errors to influence the shape of the distribution has a considerable effect on skewness when using the MPC's preferred two-piece normal density. For the 2PN, the degree of skew present in Figure 1 is lower than when outlying errors are not censored. Many of the negative forecast errors (observed during the global financial crisis) are now censored, placed in the left tail, rather than accommodated, as in Figure A1, via a higher skew estimate. The degree of skew drops further when censoring these two-years-ahead forecast errors



(a) Pre-pandemic



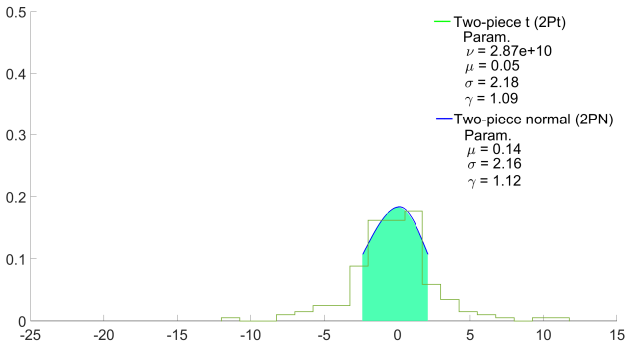
(b) Including pandemic

Figure 1: UK GDP Growth (two-years-ahead): Forecast Error Histogram and 10% Censored Two-Piece Normal and t Densities and Their Parameter Estimates Using L_A^C

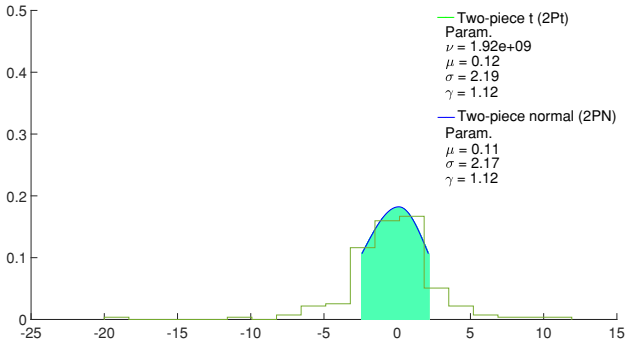
Notes: Latest release GDP estimates used to define the "outturn." Pre-pandemic: 76 outturns used from 1999q4-2018q3. Including pandemic: 84 observations used from 1999q4-2020q3. The darkest shaded green region indicates the 30% best critical region of the 2Pt; the next band extends this to 60% and the palest shade to 90%.

at 30 percent (see online Table A5). It remains the case, however, that a t distribution with a low number of degrees of freedom, 2.17, is needed to capture the peak of the error distribution.[21] Similar results are seen when forecasting one year ahead (see online Table A5). At $h = 1$ (effectively a current-quarter nowcast) the uncensored forecast error density is much more Gaussian than at the longer forecast horizons and censoring has different effects: the estimates of γ continue to be reduced by censoring, and increasingly so when censoring at 30 percent. But given that the uncensored nowcast error density is

[21]This smaller value for ν falls outside the 90 percent simulated small-sample confidence interval for a normal density; see Table A3.



(a) Pre-pandemic



(b) Simulated, including pandemic

Figure 2: US GDP Growth (one-year-ahead): FRB Forecast Error Histogram and 30% Censored Two-Piece Normal and t Densities and Their Parameter Estimates Using L_A^C

Notes: Second release GDP estimates used to define the "outturn." Pre-pandemic sample from 1974q2-2014q4. Including pandemic sample from 1974q2-2014q4 plus a single simulated observation of -20%. The darkest shaded green region indicates the 70% best critical region of the 2Pt.

effectively symmetric, this means the censored densities become right-skewed.

Once the forecast error data are updated to include the pandemic, as illustrated in Figure 1, we see a modest increase in skew for the 2Pt, but a larger increase for the 2PN. But as Table 2 shows, when symmetry is imposed and an uncensored Gaussian density is fitted, we see the error standard deviation rise due to the pandemic from 1.98 to 3.20. By contrast, when either the uncensored t density or one of the uncensored two-piece densities is fitted, the effects of the pandemic shock do not show up so obviously in increased volatility, that is, a higher standard deviation/variance estimate. In turn, if a censored rather than an uncensored Gaussian density is fitted, we again see no clear

rise in the variance of the density due to the pandemic. While there remains evidence for skew, the skew estimates are similar pre- and post-pandemic when the censored 2PN and 2Pt densities are fitted. Similar results are again seen when forecasting one year ahead. We cannot compare the nowcast errors before and after the pandemic, since the MPC did not report its usual fan chart forecast in 2020q2 when the UK economy was shut down for the first time due to COVID-19.

Turning to the US results, Figure 2 reveals that censoring (at 30 percent) eliminates the need, seen in Figure A2, for a fat-tailed density.[22] The degrees of freedom parameter rises from 3.34 in Figure A2 to, in effect, infinity in Figure 2. But there is slightly more evidence for a modest degree of left skew when censored rather than uncensored 2PN and 2Pt densities are estimated. At shorter forecast horizons, pre-pandemic (see online Table A6), the skew is to the right. But this skew is less pronounced when censoring at 10 percent rather than 30 percent. Comparing panels (a) and (b) in Figure 2, we observe few differences to the parameters of the censored 2PN and 2Pt densities before and after the pandemic. This robustness to the pandemic shock is not shared by the uncensored symmetric Gaussian density, however, as Table 2 reveals. As with the UK error data, if an uncensored Gaussian density is used, the pandemic shock shows up via a higher standard deviation. But use of a censored Gaussian density places the pandemic and the global financial crisis as outliers in the censored left tail and, accordingly, the standard deviation is similar before and including the pandemic. We also see this robustness when censoring at 10 percent rather than 30 percent (see online Table A6). The standard deviation of the censored Gaussian density is considerably smaller than its uncensored counterpart. This result applies across forecasting horizons, with the standard deviation lower still when censoring at 10 percent. The standard deviation of the censored Gaussian density (see Table 2) is a little over 2, in line with the estimates when the more fat-tailed t density or asymmetric 2PN or 2Pt densities are used, whether censored or uncensored. The censored Gaussian density also has a standard deviation similar to that of these fat-tailed and asymmetric densities at shorter forecasting horizons when continuing to censor at 30

[22]Online Table A6 shows that when censoring fewer observations, at 10 percent, a fat-tailed density is still needed. We note that for the UK, in contrast, fat-tailed densities are preferred when censoring at both 30 percent and 10 percent. Online Figure A6 also shows for the US that if an additional simulated pandemic error observation of (plus) 20 percent is appended to the error sample, to capture the rapid bounce-back of GDP in 2020q3, then while ν rises, fat tails are still required.

Table 2: Standard deviation, σ, and skew, γ, of the censored and uncensored densities fitted to the UK (MPC) and US (the FRB staff) historical forecast errors pre-pandemic and including the pandemic

height		Pre-pandemic				Incl. pandemic			
S.D.	σ	N	t	2PN	2Pt	N	t	2PN	2Pt
UK	Uncens	1.98	0.94	1.39	1.00	3.20	0.81	1.37	0.86
	Cens	1.32	0.95	1.36	0.97	1.43	0.82	1.44	0.90
US	Uncens	2.88	1.95	2.87	1.94	3.25	1.90	3.20	1.89
	Cens	2.16	2.17	2.16	2.18	2.21	2.20	2.17	2.19
		Pre-pandemic				Incl. pandemic			
Skew	γ	N	t	2PN	2Pt	N	t	2PN	2Pt
UK	Uncens	1.00	1.00	1.99	1.21	1.00	1.00	3.13	1.21
	Cens	1.00	1.00	1.25	1.14	1.00	1.00	1.49	1.22
US	Uncens	1.00	1.00	0.97	1.03	1.00	1.00	1.14	1.07
	Cens	1.00	1.00	1.12	1.09	1.00	1.00	1.12	1.12

Notes: Estimated Standard Deviation (S.D.) and skew of the censored (Cens) and uncensored (Uncens) densities fitted to two-years-ahead GDP growth forecast errors in the UK and one-year-ahead errors in the US. Reflecting institutional practice, the censored densities are at 100α=10% in the UK and 100α=30% in the US.

percent. But when fewer observations are censored, at 10 percent, evidence for fat tails (with a low estimated degrees of freedom parameter) remains. As a result, the censored Gaussian densities have higher standard deviation estimates than their t counterparts.

By way of contrast, as shown in Section A.8.2 of the online Appendix, if uncensored two-piece t and normal densities are fitted not to all the UK forecast errors as in Figure A1, but to a sub-sample of GDP growth errors before or after the global financial crisis, we also find less skew than in Figure A1. This is especially so for the two-piece normal density. This supports the view that analyzing forecast errors over a rolling window, as is apparently practiced at the Bank of England, also amounts to a form of censoring. But it is *ad hoc* and inconsistent with the fact that the density is later, at a second step, censored. In fact, when only forecast errors since the global financial crisis are used, there is no skew to the two-piece normal and the preferred density is normal (symmetric) with a variance similar to that of the two-piece t in Figure A6. One implication of this type of *censoring* is that the probability of large forecast errors is much lower than in Figure A1 or Figure 1.

In summary, we conclude that inference on the estimated parameters is affected if

the censored nature of the forecast density is acknowledged, especially when fitting, as is common, Gaussian distributions to past forecast errors. Censored error densities are, as hoped, more robust to shocks such as the global financial crisis and COVID-19 pandemic than uncensored Gaussian densities. For the UK, where, unlike in the US, there is clear evidence of asymmetry to uncensored forecast error densities perhaps explaining the MPC's decision to use the 2PN, we also see much less evidence for skew in the 2PN when censoring outlying observations. The shape of fan charts (estimated from past forecast errors) can be materially affected by whether the censoring is accommodated in estimation. As we shall see below, we also find that censoring delivers forecast error densities that exhibit fewer temporal instabilities than when an uncensored Gaussian density is used.

7 Evaluation of Censored Density Forecasts

Uncensored density forecasts are defined to be well-calibrated when they coincide with the true but unknown density. In such a case, their probability integral transforms (PITs) with respect to the outturns are uniformly distributed over the interval $(0, 1)$; for example, see Diebold, Gunther, and Tay (1998). In this section, we propose and evaluate moment-based calibration tests for multi-step-ahead censored density forecasts. These tests develop those of Knüppel (2015) and allow for the serial correlation of the PITs expected when multi-step-ahead density forecasting. We emphasize how these tests can be placed within the framework of Rossi and Sekhposyan (2019), and thereby, in the situations described therein, can accommodate parameter estimation error under the null hypothesis of correct calibration. Henceforth, we abstract from issues associated with estimation error for the parameters of the forecasting model. In our applications, the underlying models and their parameters used by the MPC and FRB are unknown (certainly to outsiders).

The point of departure when evaluating censored density forecasts is to acknowledge that while the PITs should be uniformly distributed within the uncensored region of the density, when an outturn falls in the censored region the PITs are defined only within an interval given that the censored region of the density forecast is not known probabilistically. But for a well-calibrated censored density forecast, as for an interval forecast (cf. Christoffersen (1998)), the frequency of outturns in the censored region should equal the nominal size, α. Subject to satisfying such a coverage condition, one cannot discriminate

between competing censored density forecasts in the censored region.

Let f and F continue to denote the (time-varying) probability and cumulative density functions of the (two-sided) h-step-ahead censored density forecast made at time t, and let y_{t+h} denote the subsequently observed outturn, with $t = 1, ..., T$ now denoting the out-of-sample evaluation period.

The forecast density is censored at 100α percent ($\alpha \in (0, 1)$), between the thresholds $y_{L,t}$ and $y_{U,t}$, where $y_{U,t} > y_{L,t}$, such that $\int_{y_{L,t}}^{y_{U,t}} f(y_t) = 1 - \alpha$. The PITs are defined, in the usual way, as $z_{t+h} = F(y_{t+h})$. Given the censoring, we also define PIT censoring thresholds $z_{L,t} = F(y_{L,t})$ and $z_{U,t} = F(y_{U,t})$; for example, $z_{L,t} = 0.05$ and $z_{U,t} = 0.95$ for 10 percent censoring with symmetric thresholds (about the mean). $z_{L,t}$ and $z_{U,t}$ are known by forecasters when they make their forecasts.

The censored density forecast $f(y_t)$ is well calibrated when the subset of PITs of length $T^* < T$, z^c_{t+h}, defined as those T^* elements of z_{t+h} between $z_{L,t}$ and $z_{U,t}$:

$$z^c_{t+h} = z_{t+h} \in [z_{L,t}, z_{U,t}], \tag{14}$$

are uniformly distributed between $(z_{L,t}, z_{U,t})$. So calibration involves testing $E(z^c_{t+h}) = 0.5(z_{L,t} + z_{U,t}) = 0.5$ and $Var(z^c_{t+h}) = (1/12)(z_{U,t} - z_{L,t})^2$. Outside of the uncensored range, $z_{L,t} < z_{t+h} < z_{U,t}$, calibration of $f(y_t)$ requires correct coverage:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{I}(z_{t+h} \leq z_{L,t}) + \mathbf{I}(z_{t+h} \geq z_{U,t}) = \alpha. \tag{15}$$

A joint test of uniformity, (14), and coverage, (15), is therefore required. We proceed by adding a coverage rate condition to the test of Knüppel (2015) when applied to the uncensored PITs as follows.

Let $w_t(.)$ be a threshold-weight function, such that

$$w_t(z_{t+h}) = \mathbf{I}(z_{L,t} \leq z_{t+h} \leq z_{U,t}), \tag{16}$$

where for uncensored densities, $w_t(z_{t+h}) = 1$. Let $H(.)$ denote a real-valued function of z^c_{t+h}. Berkowitz (2001) uses $H(z_{t+h}) = \Phi^{-1}(z_{t+h})$, while Knüppel (2015) considers the standardized PITs, $H(z_{t+h}) = \sqrt{12}(z_{t+h} - \frac{1}{2})$.

For censored densities, we consider a censored form of the standardized PITs:

$$\upsilon_{t+h} = H(z^c_{t+h}) = \sqrt{12/(z_{U,t} - z_{L,t})^2}\left(z^c_{t+h} - 0.5(z_{L,t} + z_{U,t})\right). \tag{17}$$

Under the null of correct calibration of the censored density between the censoring thresholds, looking at the first four moments, υ_{t+h} is uniformly distributed with an expectation of 0, variance of 1, skewness of 0, and kurtosis of 1.8, respectively.

Let $s = s_1, s_2, ..., s_N$ be a sequence of positive and finite integers, where N is the number of moment conditions under consideration ($N = 4$ here). Let $m_s = E(\upsilon_{t+h}^s)$ denote the s-th uncentered moment of υ_{t+h}, and let $\hat{m}_s = T^{*-1} \sum_{t=1}^{T^*} \upsilon_{t+h}^s$ denote its sample counterpart.

Define $\underline{\upsilon}_{t+h} = \left[\upsilon_{t+h}^{s_1}, \upsilon_{t+h}^{s_2}, ..., \upsilon_{t+h}^{s_N}\right]'$, $\underline{m} = [m_{s_1}, m_{s_2}, ..., m_{s_N}]'$ and $\underline{\hat{m}} = [\hat{m}_{s_1}, \hat{m}_{s_2}, ..., \hat{m}_{s_N}]'$. Let Ω_{T^*} denote the long-run covariance matrix of the vector:

$$\underline{d}_t = \left[\upsilon_{t+h}^{s_1} - m_{s_1}, \upsilon_{t+h}^{s_2} - m_{s_2}, ..., \upsilon_{t+h}^{s_N} - m_{s_N}\right]', \tag{18}$$

where:

$$\Omega_{T^*} = T^{*-1} \sum_{t=1}^{T^*} E\left(\underline{\upsilon}_{t+h} \underline{\upsilon}'_{t+h}\right) + T^{*-1} \sum_{j=1}^{h-1} \sum_{t=1}^{T^*} \left[E\left(\underline{\upsilon}_{t+h} \underline{\upsilon}'_{t+h-j}\right) + \sum_{t=1}^{T^*} E\left(\underline{\upsilon}_{t+h-j} \underline{\upsilon}'_{t+h}\right)\right]. \tag{19}$$

Then let $\hat{D} = [\underline{\hat{m}} - \underline{m}]'$ and $\hat{\Omega}$ be an estimator of Ω_{T^*} such that $\hat{\Omega} - \Omega_{T^*} \xrightarrow{p} 0$. Importantly, therefore, a heteroskedasticity and autocorrelation consistent (HAC) estimator can be used to estimate the covariance matrix consistently, as serial dependence is expected for correctly calibrated multi-step-ahead density forecasts.[23]

Under the null hypothesis of correct calibration of the censored density between the censoring thresholds, a moments-based test is then given as:[24]

$$\hat{\beta}_{s_1, s_2, ..., s_N} = T^* \hat{D} \hat{\Omega}^{-1} \hat{D} \xrightarrow{d} \chi_N^2. \tag{20}$$

[23] As pointed out by a referee, the properties of the HAC estimator and the ensuing moments-based tests for calibration may be affected if the autocorrelations in (18) depend on the level of ${z^c}_{t+h}$. If ${z^c}_{t+h}$ is close to the center of the interval $[z_{L,t}, z_{U,t}]$, it should have virtually the same autocorrelation as z_t. But when it is close to the interval limits, this is unlikely to be the case, because we are not going to observe ${z^c}_{t+h} = z_{L,t} - \epsilon$ or ${z^c}_{t+h} = z_{U,t} + \epsilon$ for any $\epsilon > 0$. We do not explore this issue further, but do not expect our test to suffer substantially as a result.

[24] See Knüppel (2015) for details in the uncensored case. Rossi and Sekhposyan (2019) (see their Corollary 6) generalize to maintain parameter estimation error under the null hypothesis in these raw-moment-based tests. Given how we use historical forecast error data in our application, we continue to abstract from parameter estimation error. But we note how estimation error of the parameters used to construct the densities could be preserved in our censored calibration test when, as in Rossi and Sekhposyan (2019), parameter estimation error is maintained under the null hypothesis of correct calibration. But this comes at a cost of assuming a rolling estimation window. As discussed in Section A.2 of the online Appendix and explored empirically in Section 8.2 below, rolling window estimation amounts to a form of censoring if and when outlying data fall outside the rolling estimation window.

This censored test will be compared to the regular (uncensored) Knüppel (2015) test, denoted Mom_{Uncens} below, which amounts to constructing the test statistic (20) but using the T-vector z_{t+h} rather than the T^*-vector ${z^c}_{t+h}$, where $[z_{L,t}, z_{U,t}] = [0, 1]$.

As discussed by Knüppel (2015), operationally we break (20) down into the sum of two independent test statistics calculated using the odd and even moments. While asymptotically equivalent, this test differs in small samples and was preferred by Knüppel (2015) on the basis of Monte Carlo simulations.

For censored densities we then add an additional moment condition to test coverage:

$$D_{t+h} = w_t\left(z_{t+h}\right) - (1-\alpha). \tag{21}$$

The coverage test statistic $\hat{\beta}_c$ then follows as the square of the standard normal test statistic of the sample proportion $\hat{D}_\alpha = \left(\frac{1}{T}\sum_{t=1}^{T} D_{t+h}\right)$:

$$\hat{\beta}_c = T\hat{D}_\alpha \hat{\Omega}_\alpha^{-1} \hat{D}_\alpha \xrightarrow{d} \chi^2_1, \tag{22}$$

where $\hat{\Omega}_\alpha$ is an HAC estimator of the long-run covariance of the sample proportion.[25] Like the *unconditional* coverage rate test of Christoffersen (1998), $\hat{\beta}_c$ tests the coverage of the censored density forecast but does not test whether the zeros and ones come clustered together over time. When multi-step-ahead forecasting, under correct calibration these zeros and ones need not be independent.

If interest is in testing coverage proportions in the upper and lower tails individually, rather than constructing an overall coverage test as here, then a χ^2_2 variant of $\hat{\beta}_c$ could be used instead. As Askanazi et al. (2018) explain, a consequence of using BCRs is that under correct unconditional coverage, no other set of interval forecasts (extracted from the same density forecast) with shorter intervals can also satisfy the coverage rate condition.

Finally, we construct an overall moments-based test for calibration of censored density forecasts as the sum of the two independent test statistics, (20) and (22), such that under the null of correct calibration of the censored density forecast:

$$Mom_{Cens} = \hat{\beta}_{s_1,s_2,\ldots,s_N} + \hat{\beta}_c \xrightarrow{d} \chi^2_{N+1}. \tag{23}$$

[25] Under temporal independence, required for one-step-ahead densities, this amounts to $\hat{\beta}_c = T\left(\frac{1}{T}\sum_{t=1}^{T} D_{t+h}\right)^2 / (1-\alpha)\alpha$.

Independence of the two test statistics comprising (23) arises because one test is concerned with the distribution of outturns across bins conditional on the outturns being in the uncensored region, while the other test is concerned with the distribution of outturns between the censored and uncensored regions.[26]

7.1 Monte Carlo Evidence

In this section we analyze the size and power properties of our proposed test statistic, (23). To understand performance when multi-step-ahead forecasting, we consider one-, two-, and four-step-ahead forecasts by assuming y_t is generated, respectively, by the moving average MA(0), MA(1), or MA(3) processes:

$$y_t = \varepsilon_t \tag{24}$$

$$y_t = \varepsilon_t + \rho\varepsilon_{t-1} \tag{25}$$

$$y_t = \sum_{j=0}^{3} \rho^j \varepsilon_{t-j} \tag{26}$$

where $\varepsilon_t \sim iidN(0,1)$ or $\varepsilon_t \sim iid2Pt(500,0,1,1.5)$. We set $\rho = 0.275$, a value also used by Rossi and Sekhposyan (2019).

The forecaster then constructs one-, two-, and four-step-ahead censored density forecasts of y_t from $N(0,1)$, $N(0,(1+\rho^2)^{-1})$, or $N(0,(1+\sum_{j=1}^{3}\rho^{2j})^{-1})$, respectively, with the BCR (here identical to equal-tailed) censoring intervals set as $z_{L,t} = \alpha/2$ and $z_{U,t} = (1-\alpha/2)$.

We compare the censored calibration test, (23), against its uncensored counterpart Mom_{Uncens} (namely, the four-moments test of Knüppel (2015)), the (uncensored) two-degrees-of-freedom likelihood ratio (LR) test of Berkowitz (2001) also deployed on the T-vector z_{t+h}, and a two-sided variant of the censored LR test proposed by Berkowitz (2001). This test is detailed in Appendix A.7 and, like (23), ignores the magnitude but not the frequency of forecast failure in both tails. Both of these Berkowitz-type tests are adapted, as discussed by Knüppel (2015), to the case of serially correlated PITs by

[26]Let D_i denote the event, for a particular outturn, where the PIT falls into bin i, and let A denote the event where the outturn falls in the uncensored region. The test of the PIT statistics explores the probability of $D_i|A$ while the coverage test explores the probability of A. Independence requires that P($D_i|A$).P(A)=P($D_i|A \cap$ P(A)). But this expression equals $P(D_i)$. We know that $P(D_i|A).P(A) = P(D_i \cap A)$. But further $P(D_i \cap A) = P(D_i)$ since $P(D_i \cap A') = 0$.

testing $\Phi^{-1}(z_{t+h})$ for zero mean and unit variance but not zero autocorrelation. Following Knüppel (2015), we consider the quadratic spectral kernel HAC estimator of Andrews (1991), noting that, as found in Knüppel (2015) for uncensored tests, results are similar if a Bartlett kernel as in Newey and West (1987) is used instead. [27]We analyze performance for $T = 50, 100, 250$, and $1,000$, censoring at $100\alpha = 10$ percent and $100\alpha = 30$ percent and use $10,000$ Monte Carlo simulations in each case.

7.1.1 Size analysis

To assess the size properties of the tests presented, we assume $\varepsilon_t \sim iidN(0,1)$, implying that the forecaster's Gaussian densities are correctly calibrated and deliver uniform PITs. To distinguish the censored and uncensored tests, we introduce outliers to the tails. Specifically, in the simulations when $z_{t+h} \leq z_{L,t}$ or when $z_{t+h} \geq z_{U,t}$, z_{t+h} is discarded and re-drawn from a truncated Gaussian density with mean either equal to 0 (when $z_{t+h} \leq z_{L,t}$) or 1 (when $z_{t+h} \geq z_{U,t}$), a standard deviation of 0.05, and truncation intervals defined as $[0, z_{L,t})$ for the left tail and $(z_{U,t}, 1]$ for the right tail. This implies that the PITs are uniform only between $z_{L,t}$ and $z_{U,t}$.

Panels A and B of Table 3 report the actual size of the tests at the nominal size of 5 percent when censoring at $100\alpha = 10$ percent and $100\alpha = 30$ percent. The censored test, (23), is seen to have good size properties from sample sizes of 100. Importantly, this is the case even when the PITs are serially correlated. For a sample size of 50, the test is a little under-sized. By contrast, the censored LR test of Berkowitz (2001) is under-sized even as T increases, except when the PITs are serially uncorrelated as for one-step-ahead forecasts. Even though the censored LR test is implemented - as a two degrees-of-freedom test - and so does not test for zero autocorrelation of the PITs, independence is still assumed when constructing the likelihood function.[28] As anticipated,

[27]Following the independent suggestions of a referee and Malte Knüppel, we also compared our new test against a test statistic constructed from the so-called randomized PITs; see Czado et al. (2009). This test, adapted to censored density forecasts, involves replacing (censored) values of $z_{t+h} < z_{L,t}$ with random draws from a uniform distribution defined on the interval $[0, z_{L,t}]$. Similarly, values of $z_{t+h} > z_{U,t}$ are replaced with random draws from a uniform distribution defined over the interval $[z_{U,t}, 1]$. The randomized PITs test then tests for uniformity on $[0,1]$ of these randomized PITs alongside those PITs where $z_{L,t} < z_{t+h} < z_{U,t}$. We report simulation results for this test in the online Appendix (Section A.5). The randomized test works well but Mom_{Cens} is preferred, given that unlike the randomized test, it does not assume that the censored PITs are iid.

[28]See equation (A.18) in online Appendix A.7.

both of the uncensored tests interpret nonuniformity of the tail PITs as miscalibration, and hence, when censoring at $100\alpha = 30$ percent, rejection rates rise as T increases. This feature is not seen when censoring at $100\alpha = 10$ percent, due to fewer nonuniform PITs, but we note that this feature re-emerges when T increases beyond $1,000$.

7.1.2 Power analysis

To assess power, we introduce a misspecification designed to draw out the ability of the censored calibration test to identify failures of the forecaster to characterize correctly the uncensored region of the density and the frequency of outliers. The forecaster is assumed to censor too few observations, given the outliers that affect y_t. We proceed initially as in the size experiments, with the $100\alpha = 30$ percent tail PITs randomly drawn from truncated Gaussian densities. But to test power, the forecaster is then assumed to censor only at 10 percent, meaning the censored density is now misspecified in the tails. Panel C of Table 3 shows that the censored test, (23), has good power properties for larger T. In smaller samples, power, as expected, is lower. But the power of (23) is always much higher than that of the other tests, including the censored LR test of Berkowitz (2001).

8 Production and Evaluation of Out-of-Sample MPC and FRB Censored Density Forecasts

To illustrate the real-time utility and behavior of the proposed censored density forecasts, we produce and evaluate censored density forecasts for UK and US GDP growth. The forecasts are produced as if in real time. For the UK, we also evaluate the *ex post* accuracy of the MPC's published density forecasts, correcting previous studies by acknowledging the censoring.

8.1 Production

We construct censored density forecasts by recursively fitting censored normal, t, two-piece normal (2PN), and two-piece t (2Pt) densities to the MPC and FRB historical forecast errors. These four censored densities are fitted unconditionally, both recursively adding an extra observation each quarter and using rolling windows. That is, expanding and rolling estimation windows of the historical forecast errors are used dating back to

Table 3: Size and power of censored and uncensored moment-based and LR density forecast evaluation tests

	MA0				MA1				MA3			
Panel A: Size α=0.1												
T	50	100	250	1000	50	100	250	1000	50	100	250	1000
Mom_{Uncens}	0.03	0.04	0.05	0.05	0.03	0.03	0.04	0.05	0.02	0.03	0.03	0.04
Mom_{Cens}	0.04	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.04	0.05	0.05	0.05
LR_{Uncens}	0.06	0.06	0.06	0.07	0.04	0.04	0.04	0.05	0.03	0.03	0.03	0.04
LR_{Cens}	0.05	0.05	0.05	0.05	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Panel B: Size α=0.3												
Mom_{Uncens}	0.14	0.45	0.94	1.00	0.12	0.43	0.94	1.00	0.10	0.40	0.93	1.00
Mom_{Cens}	0.03	0.04	0.04	0.05	0.02	0.04	0.04	0.05	0.02	0.04	0.04	0.05
LR_{Uncens}	0.29	0.51	0.88	1.00	0.26	0.48	0.86	1.00	0.26	0.47	0.86	1.00
LR_{Cens}	0.06	0.05	0.05	0.05	0.03	0.03	0.03	0.03	0.03	0.02	0.03	0.03
Panel C: Size-Adjusted Power												
Mom_{Uncens}	0.46	0.46	0.48	0.49	0.46	0.47	0.47	0.49	0.46	0.46	0.47	0.49
Mom_{Cens}	0.82	0.96	1.00	1.00	0.80	0.96	1.00	1.00	0.80	0.95	1.00	1.00
LR_{Uncens}	0.39	0.41	0.44	0.47	0.38	0.41	0.45	0.48	0.38	0.41	0.44	0.46
LR_{Cens}	0.71	0.86	0.99	1.00	0.71	0.86	0.98	1.00	0.72	0.86	0.98	1.00

Notes: The table reports empirical rejection frequencies for the four tests. The nominal size is 5% and the number of Monte Carlo replications is 10,000. Mom_{Uncens} is the (uncensored) Knüppel (2015) test deployed on the T-vector z_{t+h}. Mom_{Cens} is the new censored test, (23). LR_{Uncens} is the (uncensored) 2 degrees-of-freedom Berkowitz (2001) LR test. LR_{Cens} is the censored 2 degrees-of-freedom Berkowitz (2001) LR test, seen in (A.18). Panels A and B test size when censoring, respectively, at 100α=10% and 100α=30%. Panel C reports the power when the forecast density is misspecified in the tails. Power is calculated by comparing the test statistics against simulation critical values, calculated as the 95th percentile of the distributions of the statistics in the corresponding size experiment reported in Panel B.

1998q1 for the UK and to 1974q2 for the US. Following actual practice by the FOMC and the Bank of England, the rolling windows have a length of 20 years in the US and 10 years in the UK. These densities are compared with uncensored Gaussian density forecasts.

These densities are estimated using real-time GDP data vintages from the Bank of England and the Federal Reserve Bank of Philadelphia. At each forecasting origin, the outturn is defined to be the latest vintage of data, but lagged to reflect both publication lags and the fact that one has to wait one or two years to define the forecast error (for the one-year- or two-years-ahead forecasts). Note that, for the UK, the use of early rather than the latest vintage data to define the outturns against which the forecast errors are defined

represents a change from above, when we focused on the latest vintage outturns, given MPC objectives. Here, to mimic real-time use, we recursively use the latest vintage of GDP data available at the time the forecast was produced. Demonstrating the importance of GDP data revisions in the UK, we observe far more skew in the full-sample densities of the forecast errors when real-time rather than final (as of the time of writing this paper) vintage data are used; see Figures A3 and A4.

We start making forecasts for UK GDP growth in 2003q2 for 2005q1 (two years ahead), and recursively update the sample so that the final forecast we make is in 2018q4 for growth in 2020q3. For the US, the out-of-sample window is forecasts for outturns from 1994q2 to 2014q4 one year ahead.

Consistent with the earlier Monte Carlo evidence, statistical problems can arise when fitting a censored 2PN density. In particular, for the UK error data we do get divergent estimates of skew for some recursive estimates; see Figure A7 in the Appendix. The 2Pt density, however, is well behaved, and for the UK we focus on it henceforth.

In Appendix A.9 we provide graphical evidence that lets us track the evolution of the censored density forecasts over time. This shows that the censored densities exhibit more temporal stability in their moments, notably the variance, than when an uncensored Gaussian density is fitted unconditionally to a rolling window of forecast errors. This is consistent with arguments in Orlik and Veldkamp (2014): real-time estimation of densities with non-normal tails is prone to large changes in the variance - new observations "wag the tail" of the whole density. Since variance is the expected squared distance from the mean, changes in the probabilities of outliers have large effects on the conditional variance. There is less need to model time variation in forecast error standard deviations or variances when outliers are censored.

8.2 Evaluation

Table 4 evaluates the competing censored density forecasts using the censored PITs test, (23). We also separately report p-values from the two components, (20) and (22), underlying (23).

Table 4 shows that, for the UK, the data-based censored forecasts are well-calibrated with p-values greater than 0.10. While use of the rolling estimation window does, as dis-

cussed, result in some divergent estimates for the censored 2PN density, in general there is little to choose between use of the two estimation strategies. But the uncensored Gaussian density does not forecast accurately, in particular failing to get the coverage rate, condition (20), right. This is because, as seen above, the uncensored Gaussian density overstates forecast uncertainty. The better performance of the censored densities demonstrates the importance of acknowledging censoring when estimating forecast uncertainty. Use of a rolling estimation window helps improve calibration of the uncensored Gaussian density forecasts, but calibration is still worse than for any of the censored densities. These all deliver well-calibrated densities at a 90 percent significance level.

Despite the fact that the MPC can deploy judgment when forming its forecasts and, as seen, this lets them adjust to the negative shocks of the global financial crisis more rapidly, overall the MPC's forecasts when evaluated as censored densities do not appear well-calibrated at the 90 percent level. As with the Gaussian densities, their failure stems from getting the coverage rate, (20), incorrect.

For the US, Table 4 shows that all four censored densities are well-calibrated. Demonstrating the importance of accommodating temporal instabilities, the uncensored Gaussian density is well-calibrated at the 90 percent level only when a rolling estimation window is used. By contrast, the use of either an expanding or rolling estimation window again makes little difference for the censored densities. This confirms their aforementioned robustness to temporal instabilities.

9 Conclusion

This paper suggests that the uncertainty forecasts produced by a number of prominent official bodies can be interpreted as censored density forecasts. Censored density forecasts assume that some forecast uncertainties, specifically in the tails of the density, are simply not quantifiable. As in recent work that has allowed for outliers when modeling the macroeconomy in response to the COVID-19 shock (for example, see Carriero et al. (2022)), censored density forecasts require an assumption about how often outliers occur. But they have the relative attraction of not requiring an assumption about what density any outliers are then drawn from.

We examine the consequences of censoring both for the production of density forecasts

Table 4: Out-of-sample evaluation of alternative censored density forecasts for UK and US GDP growth (errors in predicting actual conditions for the period indicated): p-values from the moments-based test on the uncensored PITs test, (20), the test for coverage of the censored forecast, (22), and the joint moments-based test, (23), for the different estimators estimated using expanding and rolling windows

Window	Estimator	UK: 2005q1-2020q3			US: 1994q2-2014q4		
		(20)	(22)	(23)	(20)	(22)	(23)
Expanding	Censored 2Pt	0.12	0.18	0.10	0.21	0.79	0.31
Expanding	Censored 2PN	0.35	0.28	0.35	0.53	0.62	0.64
Expanding	Censored t	0.54	0.28	0.51	0.23	0.83	0.34
Expanding	Censored N	0.41	0.98	0.55	0.59	0.79	0.72
Expanding	Uncensored N	0.23	0.01	0.04	0.14	0.01	0.02
Rolling	Censored 2Pt	0.27	0.18	0.23	0.27	0.46	0.33
Rolling	Censored 2PN	–	–	–	0.64	0.33	0.63
Rolling	Censored t	0.39	0.10	0.24	0.33	0.45	0.39
Rolling	Censored N	0.17	0.81	0.26	0.72	0.22	0.61
Rolling	Uncensored N	0.36	0.01	0.07	0.36	0.79	0.49
	MPC	0.25	0.03	0.07			

Notes: The UK censored densities are at $\alpha = 0.1$; the US censored densities are at $\alpha = 0.3$. The length of the rolling estimation window is 10 years in the UK and 20 years in the US. "–" denotes divergent estimates with no density forecast computed.

and for their *ex post* evaluation. Accordingly, we first propose and then evaluate, through Monte Carlo, a new fixed-point algorithm that fits a potentially skewed and fat-tailed density to the inner observations but does not take a view on what distribution the outer observations come from. Our algorithm is relevant to any researcher with a small sample of data who is concerned that the outermost observations may be drawn from a distribution different from that defining the central observations. Second, we propose and evaluate a new calibration test for censored density forecasts.

We illustrate the utility of the proposed methods to produce and evaluate censored density forecasts in a context relevant for many central banks and professional forecasters. Specifically, we consider how censored density forecasts can be produced and evaluated using the historical GDP growth forecast errors made by the Federal Reserve Board staff and the MPC. Concentrating on the aftermath of the economic shocks associated with the global financial crisis and the COVID-19 pandemic, we find that the shape of the estimated density forecast is affected by whether or not one acknowledges censoring. There is less evidence for skewed and fat-tailed error densities when outer observations are censored. In other words, shocks in the tails do "wag" the whole density, as emphasized by Kozlowski, Veldkamp, and Venkateswaran (2020). Given the importance of assessments of skew for statements about the balance of risks in the macroeconomy (for example, see Adrian, Boyarchenko, and Giannone (2019)), this paper therefore demonstrates that the choice of statistical estimator used to produce the density forecast is more than a dry statistical issue. Censoring also delivers forecast error densities that exhibit fewer temporal instabilities than when an uncensored Gaussian density is used. It is an effective way of dealing with outliers that are transitory. But censoring will lead to delays in detecting regime change in the volatility and/or skew of the uncensored density, if and when the shocks are permanent. This is because shocks to the tails are censored until more than α percent are observed.

An important question for future research is whether the degree of censoring should vary over time. This could reflect the judgment of the forecaster: at times when the forecaster is especially uncertain about her probability forecasts, she may choose to censor a higher proportion. By setting the censoring at 10 percent, the MPC, for example, is stating that there is a one-in-ten chance that the unexpected will happen, although its

use of shortest-interval censored density forecasts allowing for asymmetries in the inner density means that this 10 percent need not be evenly split between the left and right tails of the forecast density. Indeed, forecasters could communicate such asymmetries directly if they wish to alert the public to upside or downside *uncertainties*, as opposed to upside and downside *risks*.

References

Adrian, Tobias, Nina Boyarchenko, and Domenico Giannone (2019). "Vulnerable growth." *American Economic Review*, 109(4), pp. 1263–1289. doi:10.1257/aer.20161923.

Alessi, Lucia, Eric Ghysels, Luca Onorante, Richard Peach, and Simon Potter (2014). "Central bank macroeconomic forecasting during the global financial crisis: The European Central Bank and Federal Reserve Bank of New York experiences." *Journal of Business and Economic Statistics*, 32(4), pp. 483–500. doi:10.1080/07350015.2014.959124.

Andrews, Donald (1991). "Heteroskedasticity and autocorrelation consistent covariance matrix estimation." *Econometrica*, 59(3), pp. 817–858. doi:10.2307/2938229.

Arellano-Valle, Reinaldo B., Héctor W. Gómez, and Fernando A. Quintana (2005). "Statistical inference for a general class of asymmetric distributions." *Journal of Statistical Planning and Inference*, 128, pp. 427–443. doi:10.1016/j.jspi.2003.11.014.

Askanazi, Ross, Francis X. Diebold, Frank Schorfheide, and Minchul Shin (2018). "On the comparison of interval forecasts." *Journal of Time Series Analysis*, 39(6), pp. 953–965. doi:10.1111/jtsa.12426.

Azzalini, Adelchi and Reinaldo B. Arellano-Valle (2013). "Maximum penalized likelihood estimation for skew-normal and skew-t distributions." *Journal of Statistical Planning and Inference*, 143, pp. 419–433. doi:10.1016/j.jspi.2012.06.022.

Berkowitz, Jeremy (2001). "Testing density forecasts with applications to risk management." *Journal of Business and Economic Statistics*, 19, pp. 465–474. doi:10.1198/07350010152596718.

Brehmer, Jonas and Tilmann Gneiting (2021). "Scoring interval forecasts: Equal-tailed, shortest, and modal interval." *Bernoulli*, 27(3), pp. 1993–2010. doi:10.3150/20-BEJ1298.

Carriero, Andrea, Todd E. Clark, Massimiliano Marcellino, and Elmar Mertens (2022).

"Addressing COVID-19 outliers in BVARs with stochastic volatility." *Review of Economics and Statistics*, Forthcoming. doi:10.1162/rest_a_01213.

Chen, Jiahua, Xianming Tan, and Runchu Zhang (2008). "Inference for normal mixtures in mean and variance." *Statistica Sinica*, 18, pp. 443–465. URL https://www.jstor.org/stable/24308490.

Christoffersen, Peter F. (1998). "Evaluating interval forecasts." *International Economic Review*, 39, pp. 841–862. doi:10.2307/2527341.

Cox, G. (2020). "Almost sure uniqueness of a global minimum without convexity." *Annals of Statistics*, 48, pp. 585–606. doi:10.1214/19-AOS1829.

Czado, Claudia, Tilmann Gneiting, and Leonhard Held (2009). "Predictive model assessment for count data." *Biometrics*, 65(4), pp. 1254–1261. doi:10.1111/j.1541-0420.2009.01191.x.

Diebold, Francis X., Todd A. Gunther, and Anthony S. Tay (1998). "Evaluating density forecasts with applications to financial risk management." *International Economic Review*, 39, pp. 863–883. doi:10.2307/2527342.

Diks, Cees, Valentyn Panchenko, and Dick van Dijk (2011). "Likelihood-based scoring rules for comparing density forecasts in tails." *Journal of Econometrics*, 163, pp. 215–230. doi:10.1016/j.jeconom.2011.04.001.

European Central Bank (2009). "New procedure for constructing eurosystem and ECB staff projection ranges." URL https://www.ecb.europa.eu/pub/pdf/other/newprocedureforprojections200912en.pdf.

Fernandez, Carmen and Mark F. J. Steel (1998). "On Bayesian modelling of fat tails and skewness." *Journal of the American Statistical Association*, 93, pp. 359–371. doi:10.1080/01621459.1998.10474117.

Gebetsberger, Manuel, Jakob W. Messner, Georg J. Mayr, and Achim Zeileis (2018). "Estimation methods for nonhomogeneous regression models: Minimum continuous ranked probability score versus maximum likelihood." *Monthly Weather Review*, 146(12), pp. 4323–4338. doi:10.1175/MWR-D-17-0364.1.

Gneiting, Tilmann and Adrian E. Raftery (2007). "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association*, 102, pp. 359–378. doi:10.1198/016214506000001437.

Haldane, Andrew G. (2012). "Tails of the unexpected." URL http://www.bankofengland.co.uk/speech/2012/tails-of-the-unexpected, speech given at "The Credit Crisis Five Years On: Unpacking the Crisis." Conference held at the University of Edinburgh Business School, June 8-9.

Hamilton, James D. (1991). "A quasi-Bayesian approach to estimating parameters for mixtures of normal distributions." *Journal of Business and Economic Statistics*, 9(1), pp. 27–39. doi:10.1080/07350015.1991.10509824.

Holzmann, Hajo and Bernhard Klar (2017). "Focusing on regions of interest in forecast evaluation." *Annals of Applied Statistics*, 11(4), pp. 2404–2431. doi:10.1214/17-AOAS1088.

Huber, Florian, Gary Koop, Luca Onorante, Michael Pfarrhofer, and Josef Schreiner (forthcoming). "Nowcasting in a pandemic using non-parametric mixed frequency VARs." *Journal of Econometrics*. doi:10.1016/j.jeconom.2020.11.006.

Hyndman, Rob J. (1996). "Computing and graphing highest density regions." *American Statistician*, 50(2), pp. 120–126. doi:10.1080/00031305.1996.10474359.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor (2020). "Disasters everywhere: The costs of business cycles reconsidered." Working Paper 26962, National Bureau of Economic Research. doi:10.3386/w26962.

Knüppel, Malte (2015). "Evaluating the calibration of multi-step-ahead density forecasts using raw moments." *Journal of Business and Economic Statistics*, 33(2), pp. 270–281. doi:10.1080/07350015.2014.948175.

Kozlowski, Julian, Laura Veldkamp, and Venky Venkateswaran (2020). "The tail that wags the economy: Beliefs and persistent stagnation." *Journal of Political Economy*, 128(8), pp. 2839–2879. doi:10.1086/707735.

Lenza, Michele and Giorgio E. Primiceri (2022). "How to estimate a vector autoregression after March 2020." *Journal of Applied Econometrics*, 37(4), pp. 688–699. doi:10.1002/jae.2895.

Newey, Whitney and Kenneth West (1987). "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix." *Econometrica*, 55(3), pp. 703–708. doi:10.2307/1913610.

Orlik, Anna and Laura Veldkamp (2014). "Understanding uncertainty shocks and the role of black swans." Working Paper 20445, National Bureau of Economic Research. doi:10.3386/w20445.

Reifschneider, David L. and Peter Tulip (2019). "Gauging the uncertainty of the economic outlook using historical forecasting errors: The Federal Reserve's approach." *International Journal of Forecasting*, 35(4), pp. 1564–1582. doi:10.1016/j.ijforecast.2018.07.016.

Rossi, Barbara and Tatevik Sekhposyan (2019). "Alternative tests for correct specification of conditional predictive densities." *Journal of Econometrics*, 208(2), pp. 638–657. doi:10.1016/j.jeconom.2018.07.008.

Sartori, Nicola (2006). "Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions." *Journal of Statistical Planning and Inference*, 136(12), pp. 4259–4275. doi:10.1016/j.jspi.2005.08.043.

Schorfheide, Frank and Dongho Song (2020). "Real-time forecasting with a (standard) mixed-frequency VAR during a pandemic." Working Paper 20-26, Federal Reserve Bank of Philadelphia. doi:10.21799/frbp.wp.2020.26.

Smith, Richard L. (1985). "Maximum likelihood estimation in a class of non-regular cases." *Biometrika*, 72, pp. 67–90. doi:10.1093/biomet/72.1.67.

Stock, James H. and Mark W. Watson (2016). "Core inflation and trend inflation." *Review of Economics and Statistics*, 98(4), pp. 770–784. doi:10.1162/REST_a_00608.

Taylor, James W. (2021). "Evaluating quantile-bounded and expectile-bounded interval forecasts." *International Journal of Forecasting*, 37(2), pp. 800–811. doi:10.1016/j.ijforecast.2020.09.007.

Tulip, Peter and Stephanie Wallace (2012). "Estimates of uncertainty around the RBA's forecasts." Research Discussion Paper 2012-07, Reserve Bank of Australia. URL https://ideas.repec.org/p/rba/rbardp/rdp2012-07.html.

Turkkan, N. and T. Pham-Gia (1997). "Highest posterior density credible region and minimum area confidence region: the bivariate case." *Journal of the Royal Statistical Society: Series C*, 46, pp. 131–140. doi:10.1111/1467-9876.00053.

Wallis, Kenneth F. (1999). "Asymmetric density forecasts of inflation and the Bank of England's fan chart." *National Institute Economic Review*, 167, pp. 106–112. doi:10.1177/002795019916700111.

White, Halbert (1980). "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica*, 48, pp. 817–838. doi:10.2307/1912934.

Woodroofe, Michael (1972). "Maximum likelihood estimation of a translation parameter of a truncated distribution." *Annals of Mathematical Statistics*, 43, pp. 113–122. doi:10.1214/aoms/1177692707.

A Online Appendix: Supplementary Material for:

Censored Density Forecasts: Production and Evaluation by Mitchell and Weale

A.1 Estimation of skew densities

In order to explore the suitability of the two-piece t and normal distributions, the focus of the main paper, we consider the general family of skew distribution parameterizations defined in Arellano-Valle, Gómez, and Quintana (2005) and Rubio and Steel (2014).[29] Like the two-piece normal, reviewed in Wallis (2014), this family of distributions involves joining two, not necessarily normal distributions with different scale parameters σ_1 and σ_2 on either side of the location parameter, μ. Specifically, Arellano-Valle, Gómez, and Quintana (2005) reparameterize these two scale parameters in terms of a common scale, σ, and a skewness parameter, α, and define the family of distributions as:

$$f(y_t|\mu,\sigma,\alpha) = \frac{2}{\sigma\left(a(\alpha)+b(\alpha)\right)} f\left(\frac{y_t-\mu}{\sigma b(\alpha)}\right) \text{ if } y_t < \mu \tag{A.1}$$

$$f(y_t|\mu,\sigma,\alpha) = \frac{2}{\sigma\left(a(\alpha)+b(\alpha)\right)} f\left(\frac{y_t-\mu}{\sigma a(\alpha)}\right) \text{ if } y_t \geq \mu \tag{A.2}$$

where f is a symmetric density and $a(\alpha)$ and $b(\alpha)$ are known and positive asymmetry functions. Asymmetries are introduced when $a(\alpha) \neq b(\alpha)$.

A leading specific density within this family (when $a(\gamma)=\gamma$, $b(\gamma)=1/\gamma$, for $\gamma>0$ and $f(.)$ is the t density) that we focus on in the main paper is the two-piece t distribution described by Fernandez and Steel (1998):[30]

$$f(y_t|\mu,\sigma,\gamma) = \frac{2}{\sigma\left(\gamma+1/\gamma\right)} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{1/2}} \left[1+\frac{(y_t-\mu)^2}{\gamma^2\nu\sigma^2}\right]^{-(\nu+1)/2} \text{ if } y_t < \mu \tag{A.3}$$

$$f(y_t|\mu,\sigma,\gamma) = \frac{2}{\sigma\left(\gamma+1/\gamma\right)} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\pi\nu)^{1/2}} \left[1+\frac{\gamma^2(y_t-\mu)^2}{\nu\sigma^2}\right]^{-(\nu+1)/2} \text{ if } y_t \geq \mu. \tag{A.4}$$

[29] Arellano-Valle, Gómez, and Quintana (2005) generalize Mudholkar and Hutson (2000), who introduced the so-called epsilon-skew-normal family of densities. This family reparameterizes the two-piece normal so that it is re-expressed in terms of an explicit skewness parameter. When this parameter equals zero, the epsilson-skew normal density reduces to the normal density.

[30] This is an instance of the so-called two-piece scale (TPSC) family of densities introduced by Rubio and Steel (2015) when $a(\alpha)=\sigma_1/\sigma$ and $b(\alpha)=\sigma_2/\sigma$, where σ_1 and σ_2 denote the scale of each of the two distributions being joined.

This estimates, as well as the location and scale parameters, the number of degrees of freedom of the t-distribution.

Generalizations of (A.1)-(A.2) involve introducing additional (shape) parameters; see Rubio and Steel (2015). Rubio and Steel's (2015) five-parameter double two-piece distribution (DTP) uses different scale but also different shape parameters on either side of the mode, μ. The DTP family contains the original two-piece densities as a subclass, as well as a four-parameter distribution (DTSH) that varies only the shape on each side of the mode. Rubio and Steel (2015) define the DTP as:

$$f(y_t|\mu,\sigma_1,\sigma_2,\delta_1,\delta_2) = \frac{2\varepsilon}{\sigma_1} f\left(\frac{y_t-\mu}{\sigma_1};\delta_1\right) \text{ if } y_t < \mu \tag{A.5}$$

$$f(y_t|\mu,\sigma_1,\sigma_2,\delta_1,\delta_2) = \frac{2(1-\varepsilon)}{\sigma_2} f\left(\frac{y_t-\mu}{\sigma_2};\delta_2\right) \text{ if } y_t \geq \mu \tag{A.6}$$

where

$$\varepsilon = \frac{\sigma_1 f\left(0;\delta_2\right)}{\sigma_1 f\left(0;\delta_2\right)+\sigma_2 f\left(0;\delta_1\right)}; \tag{A.7}$$

or

$$f(y_t|\mu,\sigma,\gamma,\delta_1,\delta_2) = \frac{2}{\sigma c(\gamma,\delta_1,\delta_2)} f\left(0;\delta_2\right) f\left(\frac{y_t-\mu}{\sigma b(\gamma)};\delta_1\right) \text{ if } y_t < \mu \tag{A.8}$$

$$f(y_t|\mu,\sigma,\gamma,\delta_1,\delta_2) = f\left(0;\delta_1\right) f\left(\frac{y_t-\mu}{\sigma a(\gamma)};\delta_2\right) \text{ if } y_t \geq \mu \tag{A.9}$$

where

$$c(\gamma,\delta_1,\delta_2) = b(\gamma) f\left(0;\delta_2\right) + a(\gamma) f\left(0;\delta_1\right). \tag{A.10}$$

Special cases of DTP include the distribution considered by Zhu and Galbraith (2010) that allows the number of degrees of freedom in (A.3)-(A.4) to be different on each side of the mode. Note also how the DTP includes four-parameter two-piece scale (TPSC) distributions, such as the two-piece t distribution seen in (A.3)-(A.4), by setting $\delta_1 = \delta_2 = \delta$, when $f(.)$ is a t density. Rubio and Steel (2015) also consider the subfamily of two-piece shape (TPSH) distributions obtained when $\sigma_1 = \sigma_2 = \sigma$ in (A.5)-(A.6). This produces distributions with different shape parameters in each direction; following Rubio and Steel (2015) let ζ explain the difference between the shapes on either side of the mode, where $\delta_1/\delta_2 = b^*(\zeta)/a^*(\zeta)$ and $\{a^*(\zeta), b^*(\zeta)\}$ are positive differentiable functions.

We consider five-parameter DTP and four-parameter DPSC and TPSH distributions with $f(.)$ chosen to be the t density and the symmetric sinh-arcsinh (SAS) distribution

of Jones and Pewsey (2009), denoted s_{JP} with asymmetry parameter ε. The SAS distribution allows for both heavier and lighter tails than the normal distribution, which is a special case when $\delta_1 = \delta_2 = 1$ and $\gamma = 0$.

As a robustness check on the results in Section 3 of the main paper, we compare the in-sample fit of the two-piece t distribution with these other classes of distributions when fitted to the MPC GDP forecast errors as seen in panel (a) of Figure A1. We consider both mature GDP outturns (as in Figure A1) and second-release GDP outturns when measuring the forecast error. As in Rubio and Steel (2015), we compare via classical information criteria (the Akaike and Bayesian information criteria: AIC and BIC) based on the ML estimates. Estimation makes use of the sn package in R (Azzalini (2018)) and R packages available at http://rpubs.com/FJRubio/DTP and http://rpubs.com/FJRubio/BTV.[31] For comparison purposes, we also consider the skew normal distribution of Azzalini (1985) and the skew t distribution of Azzalini and Capitanio (2003).[32] The skew normal of Azzalini (1985) is defined by the density function

$$f(y_t|\mu,\sigma,\alpha) = \frac{2}{\sigma}\phi(\frac{y_t-\mu}{\sigma})\Phi(\alpha\frac{y_t-\mu}{\sigma}) \tag{A.11}$$

where ϕ and Φ denote the standard normal probability density function and distribution function, respectively, and α, which regulates the skew or shape. The skew t of Azzalini and Capitanio (2003) is defined by the density function

$$f(y_t|\mu,\sigma,\alpha,\nu) = \frac{2}{\sigma}f(\frac{y_t-\mu}{\sigma}|\mu,\sigma,\nu)F(\alpha\frac{y_t-\mu}{\sigma}\sqrt{\frac{\nu+1}{\nu+\left(\frac{y_t-\mu}{\sigma}\right)^2}}|\mu,\sigma,\nu+1) \tag{A.12}$$

where f and F denote the Student t density function and distribution function, respectively, with ν degrees of freedom. Again α regulates the shape; when $\alpha = 0$ the skew t reduces to the t and when $\alpha = 0$ and $\nu = \infty$ the density reduces to the Gaussian with mean μ and standard deviation σ. And we consider the Normal Laplace distribution of Ramirez-Cobo et al. (2010), which is the convolution of a normal distribution and a two-piece Laplace distribution with location 0 and two parameters α and β. The Normal Laplace density has heavier tails than the normal density.

[31] We note that estimation of the two-piece normal and t densities in the main paper was performed in Matlab. Results were also cross-checked and verified with those from R using the sn and twopiece packages.

[32] The skew t distribution of Azzalini and Capitanio (2003) has also found recent application in macroeconomics; for example, see Adrian, Boyarchenko, and Giannone (2019).

Table A1: Pre-pandemic UK GDP two-years-ahead forecast error: 1999q4-2018q3: ML estimates of different skewed and fat-tailed density functions and AIC and BIC values

	GDP (mature data outturns)							
	AIC	BIC	$\widehat{\mu}$	$\widehat{\sigma}$	$\widehat{\gamma}$ or $\widehat{\alpha}$	$\widehat{\nu}$	$\widehat{\delta}$	$\widehat{\zeta}$
2Pt	294.57	303.90	0.03	1.01	1.30	2.51		
2PN	327.28	331.94	0.90	1.35	2.13			
DTP SAS	288.33	299.98	−1.13	15.45	−0.98		7.95	−0.96
DTP t	297.21	308.87	0.35	1.10	0.44	2.93	−0.08	
TPSC SAS	296.06	305.38	−0.16	0.52	0.22		0.52	
TPSH SAS	291.91	301.23	−0.21	0.62			0.60	−0.17
s_{JP}	293.91	303.23	−0.22	0.56	$(\widehat{\varepsilon})$ −0.22		$(\widehat{\beta})$ 0.55	
SN	304.15	311.14	1.46	3.00	−4.83			
St	293.12	302.44	0.50	1.33	−1.20	2.75		
Normal Laplace	294.58	303.90	0.65	4.76	0.62		$(\widehat{\beta})$ 0.79	
N	327.28	331.94	−0.74	2.03				
t	327.28	331.94	−0.74	2.03		15.78		
	GDP (second-release outturns)							
	AIC	BIC	$\widehat{\mu}$	$\widehat{\sigma}$	$\widehat{\gamma}$ or $\widehat{\alpha}$	$\widehat{\nu}$	$\widehat{\delta}$	$\widehat{\zeta}$
2Pt	275.51	284.84	0.54	0.67	2.50	2.93		
2PN	281.08	288.08	1.14	0.29	10.00			
DTP SAS	271.82	283.47	−0.41	9.11	−0.95		9.13	−0.95
DTP t	276.20	287.86	0.46	0.98	0.63	26.06	−0.90	
TPSC SAS	275.46	284.78	0.56	0.60	0.75		0.60	
TPSH SAS	272.78	282.10	0.01	0.74			0.85	−0.39
s_{JP}	282.32	291.65	1.77	0.02	$(\widehat{\varepsilon})$ − 6.64		$(\widehat{\beta})$ 1.22	
SN	279.90	286.89	1.15	2.93	−311573.60			
St	275.07	284.39	0.85	1.78	−6.17	3.19		
Normal Laplace	275.08	284.41	0.48	5.24	0.58		$(\widehat{\beta})$ 0.40	
N	319.97	324.63	−1.05	1.93				
t	319.97	324.63	−1.05	1.93		12.60		

Table A1 shows the ML parameter estimates and the AIC and BIC values for these 12 density functions when fitted to the two-years-ahead UK GDP growth forecast error data introduced in Section 2 and plotted in panel (a) of Figure A1. Looking across both forecast error series (defined using final and second-release GDP outturns), we see that the two-piece t fits the data competitively relative to the alternatives. While improvements in in-sample fit are achieved by the more flexible DTP and TPSC (with four or five parameters), the more parsimonious (three-parameter) two-piece t is always ranked in the top half of the 12 densities in terms of goodness of fit, according to both the AIC and the BIC. The ML parameter estimates, across the different densities, also confirm the impression from Table 2 in the main paper and Figure A1 (below) that asymmetries are important for UK GDP growth. There is also, consistent with results in the main paper, evidence that allowing for fat tails improves fit. For both sets of GDP forecast errors, the two-piece normal density (and the skew normal density of Azzalini (1985)) do not fit the data as well as the two-piece t (and the skew t density of Azzalini and Capitanio (2003)). When using second-release data, as in Figure A1, we also see that the skewed normal densities have divergent skew parameters, in contrast to the skewed t densities.

A.2 Preliminary application to MPC and FRB forecast errors: Fitting uncensored distributions

Here we illustrate the use of the uncensored 2Pt, (2), in fitting the MPC's and FRB staff's forecast errors for GDP growth. We focus attention on the MPC forecasts at the two-year horizon and the FRB's forecasts at the one-year horizon. In Section A.6 we present results for other forecasting horizons.

We show in the top panel (a) of Figure A1 the histogram of pre-pandemic (sample ends in 2018q3) forecast errors for the UK. We define the forecast error as outturn minus forecast, so negative errors are outturns below forecast. The bottom panel (b) of Figure A1 extends this sample through the pandemic to 2020q3. Figure A2 shows the forecast errors for the US. Given the publication lags associated with the *Tealbook*, for the US we have yet to observe the size of the FRB's forecast errors for the 2020 pandemic period. But it is likely that, as in the UK, especially in 2020q2 due to the lockdowns, the US will also see a large negative forecasting error. To illustrate the likely consequences for the

forecast error densities, in panel (b) of Figure A2 we therefore append to the US sample of historical forecast error data a single simulated (pandemic) error of -20 percent.

Figures A1 and A2 are full-sample histograms and use all available historical forecast error data. There are always questions about the appropriate sample over which to estimate forecasting models (and evaluate forecast accuracy). In the presence of parameter instability, due to structural breaks, there is a trade-off between bias and forecast error variance when selecting the "optimal" window of data to use. When the breaks are continuous rather than discrete, exponentially weighting observations within a window can be effective; for example, see Pesaran, Pick, and Pranovich (2013).

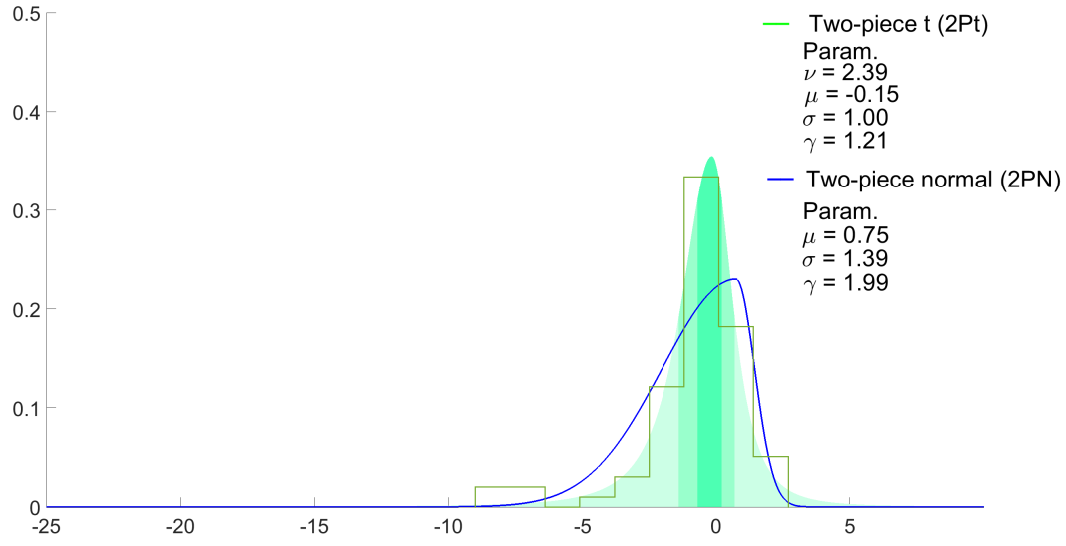
Here, given relatively small samples, we elect to use as much data as possible when estimating the unconditional densities of the forecast errors. Importantly the point of departure for our censored estimator, introduced in Section 4 in the main paper, is that it lets the whole sample (of length T) determine which specific observations within the sample to censor - for a specified coverage rate. This contrasts with our understanding of practice at the Bank of England and the FOMC. Elder et al. (2005) and Reifschneider and Tulip (2019) state that the MPC and FOMC, respectively, use rolling 10-year and 20-year windows to inform their estimates of uncertainty. At the Bank of England, in more recent years (post-financial crisis), in fact shorter windows have been used. These, in effect, censor forecast error observations not believed to be representative of current uncertainties. As well as ignoring all "old" data (certainly more than 40 quarters old) irrespective of their properties, this practice also ignores the censoring that is later imposed when the MPC publishes the fan chart only for the central 90 percent of observations.

Figures A1 and A2 also show the estimated ML parameters of the two-piece t and two-piece normal distributions fitted to the underlying forecast errors. In Figure A1, mimicking communication by the MPC, the darkest shaded region on each figure indicates the 30 percent best critical region of the two-piece t distribution. The next band extends this to 60 percent with the remaining region of the density in the palest shade. The MPC, of course, censors its density at 90 percent, an issue we turn to later. We also show on the charts the density functions estimated by fitting two-piece normal distributions. Figure A2 shows, for the US error data, the 70 percent confidence interval, as communicated by the FRB, in dark green.

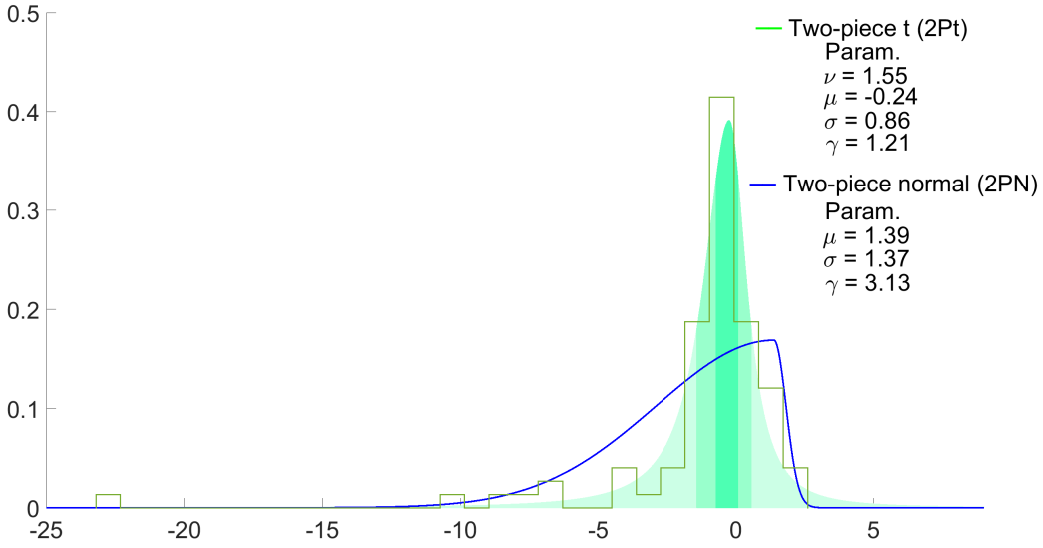
What is most striking comparing panels (a) of Figures A1 and A2 is how the forecast error densities for the UK are clearly skewed to the left, while those for the US are more symmetric, with the skew parameters for both the 2PN and 2Pt close to unity. But there is evidence for non-Gaussian features in the US, specifically fat tails, with ν estimated at 3.34. Of course, this US density is estimated on pre-pandemic forecast error data; panel (b) anticipates what effect the likely large and negative forecast error outturns will have. In the US, we see that the density remains quite symmetric, albeit with a small increase in skew. For the two-piece t we see that fatter tails are needed post-pandemic; the two-piece normal seeks to accommodate the simulated pandemic outlier observation of -20 percent via a higher standard deviation/variance estimate.

The UK error densities warrant further discussion. The left skew mentioned above is especially pronounced when the two-piece normal distribution is fitted to the GDP forecast errors. The top panel of Figure A1 shows forecast errors of up to (minus) 8 percent; these arose from a failure to forecast the global financial crisis recession of 2008-09. The number of degrees of freedom is low, at 2.39. It can be seen (statistical evaluation is provided in Section 8 of the main paper) that the two-piece t fits the center of the histogram better than the two-piece normal. This is confirmed by the aforementioned statistical evidence in Section A.1. While the t distribution is often described as having fat tails, the counterpart of this is a concentration of probability mass in the center of the distribution. The problem with the two-piece normal distribution is not so much that it means the probability of extreme events is understated, but rather that it understates the concentration of mass in the center of the distribution. It is also interesting that the two-piece t suggests less forecast error bias (a lower value for μ) than the two-piece normal density.

Comparing panels (a) and (b) of Figure A1 we isolate the effects of the pandemic. As in the US, these effects are most marked for the two-piece normal density, where the skew clearly rises to fit the large negative forecast error for 2020q2. In contrast, the parameters of the two-piece t change less, although the pandemic does result in a lower value for ν as fatter tails are needed to accommodate this historically unprecedented forecast error.



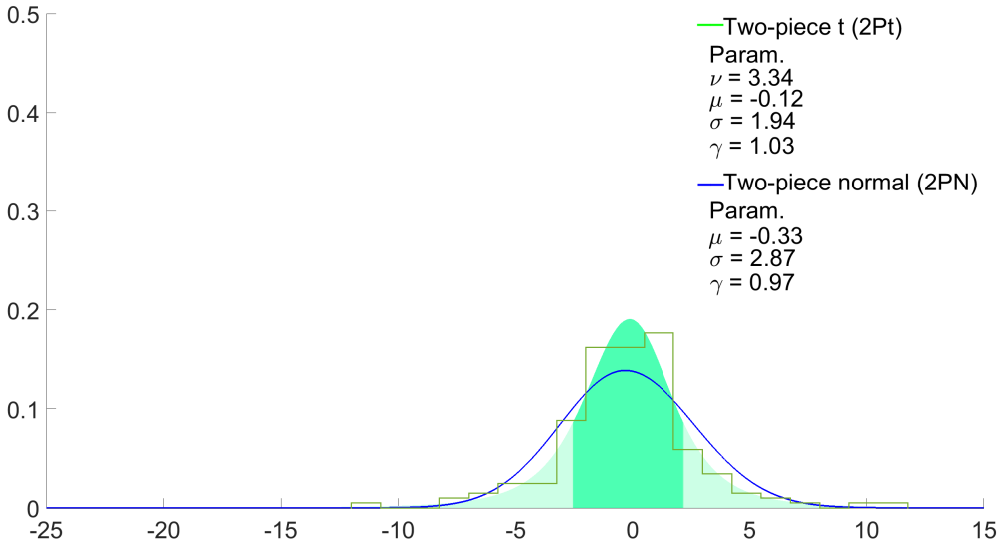
(a) Pre-pandemic



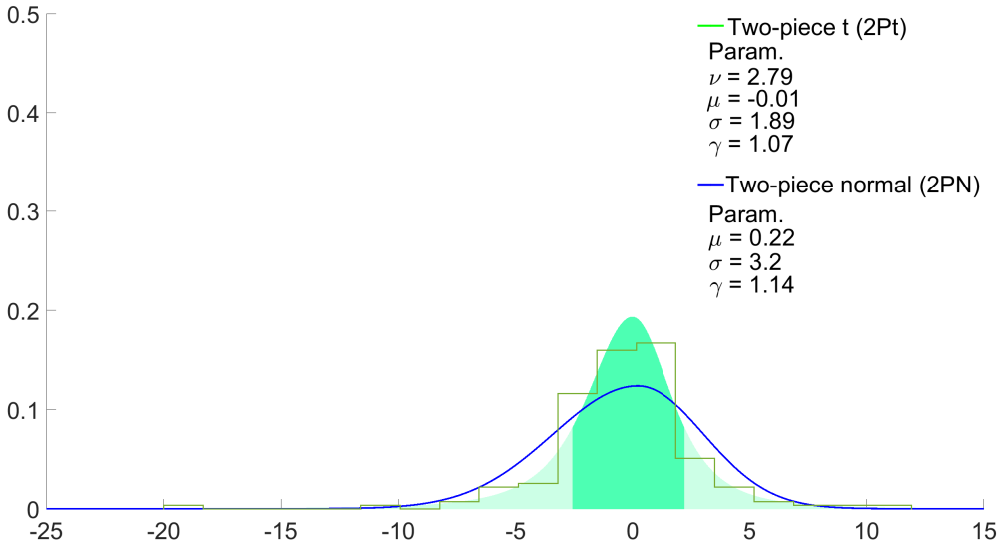
(b) Including pandemic

Figure A1: UK GDP Growth (two-years-ahead): MPC Forecast Error Histogram and Two-Piece Normal and t Densities and Their Parameter Estimates

Note: Mature GDP estimates used to define the "outturn." Pre-pandemic: 76 outturns used from 1999q4-2018q3. Including pandemic: 84 observations used from 1999q4-2020q3. The darkest shaded green region indicates the 30% best critical region of the 2Pt; the next band extends this to 60%, with the remaining region of the density in the palest shade.



(a) Pre-pandemic



(b) Simulated, including pandemic

Figure A2: US GDP Growth (one-year-ahead): FRB Forecast Error Histogram and Two-Piece Normal and t Densities and Their Parameter Estimates

Note: Second-release GDP estimates used to define the "outturn." Pre-pandemic: Sample from 1974q2-2014q4. Including pandemic: sample from 1974q2-2014q4 plus a single simulated observation of -20%. The darkest shaded green region indicates the 70% best critical region of the 2Pt.

A.3 Additional Monte Carlo results for the fixed-point estimator: Small sample confidence intervals

In this set of Monte Carlo experiments we are interested in testing whether any parameter estimates produced when fitting the censored two-piece t to the time series of forecast errors could have been, in reality, generated by an underlying symmetric normal distribution. Since our time series of pre-pandemic UK forecast errors has 76 observations, we carry out our Monte Carlo test for samples of the same length, as well as considering the smaller sample of $T = 40$ and larger samples, $T = 500$ and $1,000$. An issue we have to address is that the forecast errors relate to GDP growth over multiple quarters. If (unobserved) quarterly forecast errors are independently distributed, then errors over four quarters will follow a moving average process. If the underlying distribution is symmetric normal, then so too will be the four-quarter errors. Thus, in order to generate the data used in this experiment we draw $T + 3$ values, each denoted by u_k from a normal distribution with unit variance. We then construct T observations

$$\varepsilon_k = \left(u_k + u_{k+1} + u_{k+2} + u_{k+3}\right)/2; \;\; k = 1, ..., T \tag{A.13}$$

so that ε_k has the same variance as u_k but also follows the moving average process that arises from analysis of four-quarter forecast errors. To each set of T observations we fit the skewed t distributions, both uncensored and on the assumption that the distribution is fitted only to the central 90 percent of the observations. The true parameters values are $(1/\nu, \mu, \sigma, \gamma) = (0, 0, 1, 1)$.

Table A2: Monte Carlo results assessing the performance of the fixed-point estimator under Gaussianity: Mean, median, and standard deviation (across replications) of the parameter estimates from the censored estimators L_A^C, L_B^C, PL_A^C, and PL_B^C and the uncensored ML estimator, L; and proportion (Prop) of observations placed in the censored region plus 90% confidence intervals (CI) for the parameter estimates under normality

		T=40					**T=76**					**T=500**					**T=1,000**				
		$1/\nu$	μ	σ	γ	Prop	$1/\nu$	μ	σ	γ	Prop	$1/\nu$	μ	σ	γ	Prop	$1/\nu$	μ	σ	γ	Prop
	true	**0.00**	**0.00**	**1.00**	**1.00**		**0.00**	**0.00**	**1.00**	**1.00**		**0.00**	**0.00**	**1.00**	**1.00**		**0.00**	**0.00**	**1.00**	**1.00**	
L_A^C	mean	0.09	-0.01	0.75	3.12	0.09	0.06	0.01	0.87	1.11	0.09	0.03	0.01	0.97	1.01	0.10	0.03	0.00	0.98	1.00	0.10
	med	0.00	0.00	0.76	0.99	0.08	0.00	-0.01	0.87	1.01	0.09	0.00	-0.01	0.97	1.00	0.10	0.00	0.00	0.98	1.00	0.10
	sd	0.17	0.71	0.28	17.77	0.03	0.12	0.53	0.17	1.61	0.02	0.05	0.19	0.07	0.11	0.00	0.04	0.13	0.05	0.07	0.00
	low CI	**0.00**	**-1.20**	**0.04**	**0.32**	**0.03**	**0.00**	**-0.87**	**0.60**	**0.58**	**0.07**	**0.00**	**-0.29**	**0.85**	**0.85**	**0.09**	**0.00**	**-0.20**	**0.89**	**0.89**	**0.09**
	up CI	**0.40**	**0.88**	**1.09**	**1.97**	**0.14**	**0.37**	**0.68**	**1.10**	**1.48**	**0.11**	**0.11**	**0.38**	**1.06**	**1.24**	**0.10**	**0.10**	**0.22**	**1.05**	**1.12**	**0.10**
L_B^C	mean	0.08	0.00	0.50	42.77	0.08	0.06	0.00	0.69	9.98	0.09	0.03	0.02	0.95	1.04	0.10	0.03	0.00	0.97	1.01	0.10
	med	0.00	0.01	0.54	0.97	0.08	0.00	0.02	0.78	1.02	0.09	0.00	0.00	0.96	1.02	0.10	0.00	0.00	0.97	1.00	0.10
	sd	0.19	0.86	0.45	226.3	0.04	0.14	0.71	0.34	53.84	0.02	0.05	0.25	0.07	0.22	0.01	0.04	0.17	0.05	0.14	0.00
	low CI	**0.00**	**-1.41**	**0.00**	**0.00**	**0.03**	**0.00**	**-1.24**	**0.00**	**0.00**	**0.05**	**0.00**	**-0.39**	**0.83**	**0.73**	**0.09**	**0.00**	**-0.28**	**0.88**	**0.80**	**0.09**
	up CI	**0.37**	**1.10**	**1.01**	**101.2**	**0.14**	**0.44**	**1.12**	**1.03**	**5.09**	**0.12**	**0.13**	**0.37**	**1.07**	**1.32**	**0.10**	**0.11**	**0.26**	**1.04**	**1.19**	**0.10**
PL_A^C	mean	0.05	0.01	0.78	2.41	0.09	0.06	-0.04	0.88	1.02	0.09	0.04	-0.03	0.97	0.99	0.10	0.03	-0.02	0.98	0.99	0.10
	med	0.00	0.05	0.80	1.03	0.10	0.00	-0.04	0.88	0.99	0.09	0.00	-0.03	0.97	0.98	0.10	0.00	-0.02	0.98	0.98	0.10
	sd	0.12	0.73	0.26	12.88	0.03	0.11	0.48	0.16	0.32	0.02	0.05	0.17	0.07	0.10	0.01	0.04	0.14	0.05	0.08	0.00
	low CI	**0.00**	**-1.28**	**0.14**	**0.46**	**0.05**	**0.00**	**-0.78**	**0.62**	**0.59**	**0.06**	**0.00**	**-0.30**	**0.86**	**0.83**	**0.09**	**0.00**	**-0.25**	**0.89**	**0.86**	**0.09**
	up CI	**0.30**	**1.12**	**1.13**	**2.37**	**0.13**	**0.29**	**0.74**	**1.15**	**1.56**	**0.12**	**0.15**	**0.22**	**1.08**	**1.17**	**0.10**	**0.10**	**0.22**	**1.05**	**1.13**	**0.10**
PL_B^C	mean	0.06	0.03	0.55	3.56	0.09	0.04	-0.08	0.73	1.35	0.09	0.03	-0.06	0.95	0.97	0.10	0.02	-0.05	0.97	0.97	0.10
	med	0.00	0.09	0.65	1.03	0.10	0.00	-0.08	0.80	0.97	0.09	0.00	-0.07	0.95	0.94	0.10	0.00	-0.05	0.98	0.95	0.10
	sd	0.15	0.89	0.37	14.89	0.03	0.11	0.67	0.31	2.27	0.02	0.05	0.25	0.08	0.21	0.01	0.04	0.17	0.05	0.13	0.00
	low CI	**0.00**	**-1.47**	**-0.01**	**0.00**	**0.05**	**0.00**	**-1.24**	**0.00**	**0.00**	**0.06**	**0.00**	**-0.46**	**0.83**	**0.70**	**0.09**	**0.00**	**-0.31**	**0.87**	**0.77**	**0.09**
	up CI	**0.43**	**1.41**	**1.04**	**7.89**	**0.13**	**0.32**	**1.02**	**1.11**	**3.61**	**0.12**	**0.14**	**0.43**	**1.07**	**1.31**	**0.11**	**0.11**	**0.26**	**1.04**	**1.22**	**0.10**
L	mean	0.03	0.04	0.83	1.13		0.02	0.03	0.91	1.05		0.01	0.04	0.98	1.03		0.01	0.02	0.98	1.02	
	med	0.00	0.12	0.83	1.03		0.00	0.12	0.90	1.03		0.00	0.07	0.98	1.03		0.00	0.04	0.98	1.02	
	sd	0.07	0.51	0.23	1.07		0.05	0.39	0.15	0.26		0.02	0.16	0.06	0.09		0.02	0.12	0.04	0.07	
	low CI	**0.00**	**-0.95**	**0.49**	**0.52**		**0.00**	**-0.70**	**0.67**	**0.65**		**0.00**	**-0.26**	**0.88**	**0.87**		**0.00**	**-0.19**	**0.91**	**0.90**	
	up CI	**0.17**	**0.69**	**1.11**	**1.57**		**0.14**	**0.46**	**1.10**	**1.37**		**0.07**	**0.30**	**1.07**	**1.18**		**0.04**	**0.18**	**1.04**	**1.11**	

Table A2 shows the mean, median, and standard deviation of each parameter taken from the $R = 1,000$ draws, together with the upper and lower 90 percent confidence limits. These are calculated by taking the 50th and 950th values of the ordered parameter estimates across the 1,000 Monte Carlo replications.

A number of things stand out from Table A2, beyond the obvious point that the fit of L_A^C and L_B^C is much worse, with the confidence intervals wider, with small samples than with $T = 1000$ observations. $1/\nu$ cannot be expected to be symmetric around its true value of zero, so a bias inevitably exists in the small-sample estimates. A related bias appears in the estimate of σ. A low number of degrees of freedom and a high value of σ are both ways of accommodating observations distant from the mode, so bias in one implies a bias in the other. There is little evidence of bias in the mode, μ, since the confidence limits are reasonably symmetric. γ appears skewed to the right, especially so for small samples and for L_B^C rather than L_A^C; the confidence limits are asymmetric.

When the distribution is not censored the estimates for σ and γ in Table A2 are better determined than when censor points are estimated simultaneously. This is not very surprising. But for smaller samples a bias does appear in L's estimates for μ. It is also worth noting that, even when the distribution is not censored, when $T = 76$ an estimate of $1/0.14 = 7$ degrees of freedom has a 5 percent chance of arising from an underlying normal distribution.

When the data are censored so that the distribution is fitted to only the central 90 percent of observations, the estimated value of the number of degrees of freedom has to be 2.7 (2.2) or lower under L_A^C (L_B^C) before one can reject, at a 90 percent significance level, the hypothesis that the underlying distribution is normal.

There is again evidence of a higher possibility of divergence in the censored estimates of γ (and in turn those for σ) for smaller samples as evidenced by a higher standard deviation for γ. But L_A^C is less contaminated than L_B^C by some high values for γ. Contamination for both estimators is worse when T drops from 76 to 40, which should be borne in mind in our out-of-sample application (Section 8 of the main paper). We note that the median estimates for γ from L_A^C and L_B^C remain accurate, close to unity, even when $T = 40$. But, as seen from comparison with Table 1 (in the main paper), this feature is specific to the situation when there is no skew in the population data. Recall we found that when there

is population data skew, the median estimates for γ from L_A^C and L_B^C differ from the true value - and the penalized estimators are likely to be preferred in such small samples. Table A2 shows that with symmetric data the penalized estimators continue to lower the chance of divergent estimates for γ.

Overall, Table A2 shows that in small samples L_A^C continues to be preferred to L_B^C as its estimates for the four parameters are better determined: its mean and median estimates are closer to the true values with lower standard deviations and tighter and more symmetric confidence bands. But (even without the population data skew considered in Table 1 in the main paper, and Table A3) there remains a chance in smaller samples that estimation using L_A^C delivers divergent values for γ. So in practice, including in the out-of-sample application in Section 8, we recommend looking closely at the parameter estimates for fear they involve an (economically) unappealing boundary value for γ. If the estimated value of γ diverges, the resulting density in effect becomes a half or folded density; for an illustration of such a cliff-edged density see Figure A3 (in this online Appendix). If estimates do diverge, based on the results in Table 1 and Table A3, we suggest use of our penalized estimator as it is found to mitigate, albeit not eradicate, the possibility of boundary values in small samples. Moreover, in any applications when estimation does appear to reflect divergence - with the estimates of γ (or $1/\gamma$) rising above a threshold value of say 5 or 10 implying a half or folded density - the estimates might be rejected and model/density estimation reconsidered.

A.4 Additional Monte Carlo results for the fixed-point estimator: Performance for different sample sizes

This set of simulations tests the performance of the censored estimator, under both L_A^C and L_B^C, in samples of different sizes as the degree of skew varies. Comparison is made with the uncensored ML estimator, L. T observations are drawn from a two-piece t distribution, where $(\nu, \mu, \sigma, \gamma) = (5, 0, 1, 1.5)$ and $(5, 0, 1, 2.5)$. $\gamma = 1.5$ and $\gamma = 2.5$ correspond to moderate and high (positive) skew. We consider $T = 40, 100, 500$, and $1,000$, noting the Bank of England's use of just 40 observations to estimate its forecast error densities. We focus on censoring at 100α=10 percent, also in keeping with MPC practice. We report results based on $R = 1,000$ replications. For $\gamma = 2.5$, we also report results for $PL_j^C(y_t, \beta)$.

We do not report results for $PL_j^C(y_t, \beta)$ when $\gamma = 1.5$, since, as will be seen, the utility of the penalized estimators, relative to the unpenalized ones, is found to be greater in populations with high skew.

Table A3: Monte Carlo results assessing the performance of the fixed-point estimator for different sample sizes: Mean, median, and standard deviation (across replications) of the parameter estimates from the censored estimators L_A^C, L_B^C, PL_A^C, and PL_B^C and the uncensored ML estimator, L; and proportion (Prop) of observations placed in the censored region

		T=40					**T=100**					**T=500**					**T=1,000**				
		$1/\nu$	μ	σ	γ		$1/\nu$	μ	σ	γ		$1/\nu$	μ	σ	γ		$1/\nu$	μ	σ	γ	
true		**0.20**	**0.00**	**1.00**	**1.50**	**Prop**	**0.20**	**0.00**	**1.00**	**1.50**	**Prop**	**0.20**	**0.00**	**1.00**	**1.50**	**Prop**	**0.20**	**0.00**	**1.00**	**1.50**	**Prop**
L_A^C	**mean**	0.21	0.01	0.90	2E+4	0.10	0.19	0.02	0.98	1.58	0.10	0.20	0.00	1.00	1.51	0.10	0.20	0.00	1.00	1.50	0.10
	med	0.15	0.03	0.91	1.55	0.10	0.18	0.02	0.98	1.52	0.10	0.20	0.00	1.00	1.51	0.10	0.20	0.00	1.00	1.50	0.10
	sd	0.23	0.52	0.31	5E+5	0.02	0.16	0.29	0.16	0.35	0.01	0.08	0.11	0.07	0.11	0.00	0.06	0.08	0.05	0.08	0.00
L_B^C	**mean**	0.19	0.01	0.63	7E+4	0.10	0.17	0.08	0.84	63.39	0.10	0.20	0.01	0.98	1.56	0.10	0.20	0.00	0.99	1.52	0.10
	med	0.00	0.12	0.70	1.81	0.10	0.12	0.07	0.91	1.62	0.10	0.20	0.01	0.98	1.53	0.10	0.20	0.00	0.99	1.50	0.10
	sd	0.26	0.79	0.72	2E+6	0.06	0.18	0.51	0.34	834	0.05	0.08	0.18	0.10	0.27	0.00	0.06	0.12	0.07	0.17	0.00
L	**mean**	0.16	0.03	0.95	3.09		0.17	0.02	1.00	1.58		0.20	0.01	1.00	1.51		0.20	0.00	1.00	1.50	
	med	0.13	0.07	0.97	1.54		0.17	0.03	0.99	1.52		0.20	0.01	1.00	1.51		0.20	0.00	1.00	1.50	
	sd	0.16	0.45	0.27	9.90		0.11	0.27	0.15	0.35		0.05	0.11	0.06	0.11		0.03	0.08	0.04	0.07	
true		**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**
L_A^C	**mean**	0.20	0.02	0.71	1E+5	0.09	0.19	0.05	0.88	126.3	0.10	0.19	0.01	0.99	2.84	0.10	0.20	0.00	1.00	2.52	0.10
	med	0.09	0.10	0.78	2.83	0.10	0.17	0.06	0.92	2.65	0.10	0.19	0.00	0.99	2.51	0.10	0.20	0.00	1.00	2.50	0.10
	sd	0.24	0.53	0.52	2E+6	0.05	0.17	0.32	0.34	858.9	0.02	0.08	0.14	0.12	6.98	0.01	0.06	0.09	0.07	0.19	0.00
L_B^C	**mean**	0.16	0.05	0.41	2E+5	0.12	0.16	0.12	0.58	3E+3	0.11	0.19	0.03	0.91	33.9	0.10	0.19	0.01	0.98	5.08	0.10
	med	0.00	0.18	0.04	61.04	0.10	0.00	0.21	0.57	4.75	0.10	0.19	0.02	0.98	2.56	0.10	0.20	0.01	0.99	2.51	0.10
	sd	0.25	1.27	0.53	4E+6	0.17	0.19	0.46	0.52	6E+4	0.07	0.10	0.23	0.31	225	0.01	0.07	0.14	0.16	38.1	0.00
L	**mean**	0.15	0.06	0.79	16.31	0.00	0.17	0.04	0.92	5.24		0.19	0.00	0.99	2.55		0.20	0.00	1.00	2.52	
	med	0.11	0.10	0.88	2.81	0.00	0.17	0.06	0.96	2.68		0.19	0.01	0.99	2.53		0.20	0.00	1.00	2.50	
	sd	0.15	0.41	0.45	28.36	0.00	0.10	0.27	0.27	12.50		0.04	0.11	0.09	0.25		0.03	0.08	0.06	0.17	
true		**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**	**0.20**	**0.00**	**1.00**	**2.50**	**Prop**
PL_A^C	**mean**	0.18	0.01	0.76	1385	0.10	0.17	-0.02	0.97	193.4	0.10	0.18	-0.05	1.03	2.44	0.10	0.19	-0.04	1.03	2.42	0.10
	med	0.08	0.07	0.82	2.79	0.10	0.15	0.00	0.99	2.50	0.10	0.18	-0.04	1.03	2.41	0.10	0.19	-0.04	1.03	2.41	0.10
	sd	0.23	0.51	0.47	2E+4	0.03	0.16	0.30	0.28	5E+3	0.02	0.08	0.13	0.11	0.26	0.00	0.06	0.10	0.07	0.19	0.00
PL_B^C	**mean**	0.13	-0.10	0.77	7E+5	0.11	0.11	-0.10	0.98	167.3	0.11	0.15	-0.12	1.10	2.70	0.11	0.16	-0.12	1.11	2.29	0.00
	med	0.00	0.10	0.74	3.29	0.10	0.00	-0.02	1.02	2.43	0.11	0.17	-0.10	1.10	2.22	0.10	0.19	-0.10	1.09	2.23	0.00
	sd	0.22	0.71	0.49	2E+7	0.03	0.17	0.46	0.42	3978	0.02	0.12	0.24	0.22	7.25	0.01	0.10	0.18	0.18	0.46	0.00

The mean and median values and the standard deviations (across the $1,000$ replications) of the estimates of the four parameters are shown in Table A3. We also report the proportion (averaged across the R replications) of the T observations that, for the censored estimators, are classified as falling in the censored region.

Not surprisingly, for large samples ($T = 1,000$) Table A3 shows higher standard errors for the parameters fitted to the censored data using the fixed-point method than to the uncensored data by ML. At the same time, however, the results confirm that in large samples the censored estimators work well, especially when skew is moderate ($\gamma = 1.5$) rather than extreme ($\gamma = 2.5$): under both L_A^C and L_B^C when $\gamma = 1.5$ the mean and median values equal (to two decimal places) those in the data-generating process.[33] When $\gamma = 2.5$, L_A^C continues to have this property but there is extra noise in the estimates for L_B^C, which imposes less structure than L_A^C. This is seen by L_B^C showing deviations from the true parameter estimates for γ. These deviations reflect a few outlying estimates, for some iterations, with the median values closer to the true parameter values than the mean ones. Both L_A^C and L_B^C correctly place, on average across R, 10 percent of observations in the censored region with little variation even for small T.

Once the sample size drops to 100, problems with estimation of γ start to appear. This is so for L_A^C but especially L_B^C. An increasing number of the R draws return inaccurate (high, divergent) estimates of γ: the median parameter estimates remain closer to the true values than the mean ones.[34] With L_A^C there is slight evidence of bias (looking at the mean across replications) when the true value is $\gamma = 1.5$; but when L_B^C is used instead, we can see that the mean parameter estimate for γ is contaminated by some very high values (for these draws this is accompanied by extremely low values for σ). Essentially the divergence problems reported by Sartori (2006) and Azzalini and Arellano-Valle (2013) emerge. The problem is worse when $\gamma = 2.5$ than when $\gamma = 1.5$, and there is evidence of it even when the distribution is not censored and the likelihood function L is used. These problems become more acute when the sample drops to 40 observations. At this stage, the mean estimates for γ from L_A^C, L_B^C and L all give contaminated results due

[33] Convergence was also satisfied, with P_r converging to zero. An alternative and simpler method that would also work in large samples would be to set fixed censor points to exclude the upper and lower 10 percent of observations. This would allow the parameters to be estimated straightforwardly.

[34] We note that if we were to assume negative rather than positive skew in the data-generating process, that is, $\gamma = 1/\gamma$, then the estimated γ are at risk of diverging to zero rather than infinity.

to the increased risk that for some replications the estimates for γ diverge. Use of the penalized estimator does help in these smaller samples, when $\gamma = 2.5$ especially for L_B^C. While it does not prevent the mean estimate for γ (across replications) from rising above the true value, the median estimates are closer to the true values than when a penalty is not imposed. We therefore conclude that in very small samples it may prove helpful, in effect, to have a prior that the data are symmetric. But even if they are not, imposing this view via the penalized estimator improves the accuracy of the median estimates even when the data are in fact highly skewed. The penalized censored estimators continue to place 10 percent of the 40 observations in the censored region. In larger samples (for example, $T = 1,000$), imposing a penalty does cause γ to be underestimated slightly, and in turn σ to be overestimated. But this bias is relatively modest, about 1 percent for L_A^C (for the median estimates) and about double this for L_B^C.

A.5 Additional Monte Carlo results for the tests on the evaluation of censored density forecasts

Here we supplement Table 3 in the main paper with the results from the randomization test (Rand) discussed in footnote 27 of the main paper.

The randomization test is implemented by replacing (censored) values of $z_{t+h} < z_{L,t}$ with random draws from a uniform distribution defined on the interval $[0, z_{L,t}]$. Similarly, values of $z_{t+h} > z_{U,t}$ are replaced with random draws from a uniform distribution defined over the interval $[z_{U,t}, 1]$. The randomized PITs test then tests for uniformity on $[0, 1]$ of these randomized PITs alongside those PITs where $z_{L,t} < z_{t+h} < z_{U,t}$. Specifically, we test uniformity of this augmented PITs vector using the four-moments-based test of Knüppel (2015), as discussed in the main paper. Note how this test does therefore allow for serial correlation in the uncensored PITs. But, as summarized in the main paper, the randomization test draws the censored PITs under the assumption of iid-ness. It is this assumption that helps to explain why the size and power properties of the randomization test (seen in Table A4) are somewhat worse when multi-step-ahead forecasting than those for Mom_{Cens}.

Table A4: Size and power of censored and uncensored moment-based, LR density forecast, and randomization evaluation tests

	MA0				MA1				MA3			
T	50	100	250	1000	50	100	250	1000	50	100	250	1000
Panel A: Size α=0.1												
Mom_{Uncens}	0.03	0.04	0.05	0.05	0.03	0.04	0.04	0.05	0.02	0.03	0.04	0.04
Mom_{Cens}	0.04	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.04	0.05	0.05	0.05
LR_{Uncens}	0.06	0.06	0.06	0.07	0.04	0.04	0.03	0.05	0.03	0.03	0.03	0.04
LR_{Cens}	0.05	0.05	0.05	0.05	0.04	0.04	0.03	0.03	0.03	0.03	0.03	0.03
Rand	0.03	0.04	0.04	0.05	0.02	0.03	0.04	0.04	0.02	0.03	0.03	0.04
Panel B: Size α=0.3												
Mom_{Uncens}	0.16	0.45	0.94	1.00	0.12	0.43	0.93	1.00	0.11	0.41	0.93	1.00
Mom_{Cens}	0.03	0.04	0.04	0.05	0.02	0.04	0.04	0.05	0.02	0.04	0.04	0.05
LR_{Uncens}	0.29	0.51	0.88	1.00	0.25	0.47	0.86	1.00	0.25	0.46	0.86	1.00
LR_{Cens}	0.06	0.05	0.05	0.05	0.03	0.03	0.03	0.03	0.02	0.03	0.02	0.03
Rand	0.03	0.04	0.05	0.05	0.02	0.03	0.04	0.05	0.02	0.03	0.03	0.04
Panel C: Size-Adjusted Power												
Mom_{Uncens}	0.46	0.46	0.48	0.49	0.46	0.47	0.47	0.49	0.46	0.46	0.47	0.49
Mom_{Cens}	0.82	0.96	1.00	1.00	0.80	0.96	1.00	1.00	0.80	0.95	1.00	1.00
LR_{Uncens}	0.39	0.41	0.44	0.47	0.38	0.41	0.45	0.48	0.38	0.41	0.44	0.46
LR_{Cens}	0.71	0.86	0.99	1.00	0.71	0.86	0.98	1.00	0.72	0.86	0.98	1.00
Rand	0.74	0.91	1.00	1.00	0.74	0.90	1.00	1.00	0.74	0.91	0.99	1.00

Notes: See notes to Table 3 in the main paper. Rand is the randomization test. The table reports empirical rejection frequencies for the four tests of Table 3 and the randomization test. The nominal size is 5% and the number of Monte Carlo replications is 10,000. Panels A and B test size when censoring, respectively, at 100α=10% and 100α=30% . Panel C reports the size-adjusted power when the forecast density is misspecified in the tails.

A.6 Censored densities fitted to MPC and FRB GDP forecast errors: Cross-horizon and censoring at 10 percent and 30 percent results

Tables A5 and A6 report the estimated parameters of the censored and uncensored densities fitted to the UK and US forecast error data at various forecasting horizons, pre-pandemic and including the pandemic with censoring at 100α=10 percent and 100α=30 percent. These results are referenced in Section 6 of the main paper.

Table A5: Standard deviation, σ, skew, γ, and degrees of freedom, ν, of the censored and uncensored densities fitted to the UK historical GDP forecast errors pre-pandemic and including the pandemic: estimated parameters at different forecasting horizons (in quarters), h, with censoring at α=0.1 and α=0.3

		Pre-pandemic				Incl. pandemic			
		N	t	2PN	2Pt	N	t	2PN	2Pt
S.D.	σ								
$h=1$	Uncens	1.07	1.07	1.07	1.07	-	-	-	-
	Cens 70	1.03	0.83	0.89	0.76	-	-	-	-
	Cens 90	1.08	1	1.06	1.06	-	-	-	-
$h=5$	Uncens	1.77	0.93	1.57	0.93	3.06	0.81	1.79	0.82
	Cens 70	1.12	0.99	1.21	0.94	1.16	0.86	1.14	0.86
	Cens 90	1.26	0.95	1.26	0.93	1.28	0.81	1.36	0.77
$h=8$	Uncens	1.98	0.94	1.39	1.00	3.20	0.81	1.37	0.86
	Cens 70	1.17	0.84	1.21	0.79	1.20	0.76	1.23	0.76
	Cens 90	1.32	0.95	1.36	0.97	1.43	0.82	1.44	0.90
Skew	γ								
$h=1$	Uncens	1.00	1.00	0.97	1.08	-	-	-	-
	Cens 70	1.00	1.00	0.64	0.66	-	-	-	-
	Cens 90	1.00	1.00	0.86	0.86	-	-	-	-
$h=5$	Uncens	1.00	1.00	1.42	1.11	1.00	1.00	2.25	1.18
	Cens 70	1.00	1.00	1.14	1.03	1	1	1.16	1.08
	Cens 90	1.00	1.00	0.99	1.05	1.00	1.00	1.16	1.14
$h=8$	Uncens	1.00	1.00	1.99	1.21	1.00	1.00	3.13	1.21
	Cens 70	1.00	1.00	0.90	0.87	1.00	1.00	1.01	0.99
	Cens 90	1.00	1.00	1.25	1.14	1.00	1.00	1.49	1.22
dof	ν								
$h=1$	Uncens	-	1000	-	1000	-	-	-	-
	Cens 70	-	2.42	-	3.03	-	-	-	-
	Cens 90	-	9.26	-	1000	-	-	-	-
$h=5$	Uncens	-	2.26	-	2.33	-	1.51	-	1.58
	Cens 70	-	3.48	-	2.29	-	1.80	-	1.88
	Cens 90	-	2.63	-	2.37	-	1.50	-	1.30
$h=8$	Uncens	-	2.01		2.39	-	1.39	-	1.55
	Cens 70	-	1.40		1.22	-	1.20	-	1.18
	Cens 90	-	2.10		2.17	-	1.37	-	1.78

Notes: Estimated parameters of the 100α=10% (Cens 90) and 100α=30% (Cens 70) and uncensored (Uncens) densities. Post-pandemic estimates are not available at $h=1$ as the MPC chose not to report its fan chart forecasts in 2020q2. Pre-pandemic sample (with dates referring to the outturn) is from 1998q1-2019q4 at $h=1$, 1999q1-2018q3 at $h=5$ and 1999q4-2018q3 at $h=8$. Incl. pandemic sample (with dates referring to the outturn) is from 1998q1-2019q4 at $h=1$, 1999q1-2020q3 at $h=5$ and 1999q4-2020q3 at $h=8$. Latest vintage GDP estimates used to define the GDP growth outturns.

Table A6: Standard deviation, σ, skew, γ, and degrees of freedom, ν, of the censored and uncensored densities fitted to the US historical GDP forecast errors pre-pandemic and including the pandemic: estimated parameters at different forecasting horizons (in quarters), h, and censoring at α=0.1 and α=0.3

		Pre-pandemic				Incl. pandemic			
		N	t	2PN	2Pt	N	t	2PN	2Pt
S.D.	σ								
$h=1$	Uncens	2.17	1.91	2.11	1.83	2.27	1.73	2.57	1.64
	Cens 70	1.91	1.85	1.69	1.72	1.87	1.87	1.74	1.74
	Cens 90	2.14	1.59	2.07	1.51	2.17	1.64	2.14	1.52
$h=3$	Uncens	2.98	2.27	2.96	2.27	3.34	2.18	3.17	2.19
	Cens 70	2.46	2.48	2.49	2.21	2.47	2.24	2.53	2.24
	Cens 90	2.66	2.24	2.66	2.23	2.72	2.15	2.69	2.13
$h=5$	Uncens	2.88	1.95	2.87	1.94	3.25	1.90	3.20	1.89
	Cens 70	2.16	2.17	2.16	2.17	2.21	2.20	2.20	2.18
	Cens 90	2.45	1.90	2.46	1.86	2.47	1.88	2.51	1.89
Skew	γ								
$h=1$	Uncens	1.00	1.00	0.82	0.81	1.00	1.00	1.23	0.83
	Cens 70	1.00	1.00	0.66	0.70	1.00	1.00	0.73	0.73
	Cens 90	1.00	1.00	0.78	0.77	1.00	1.00	0.82	0.80
$h=3$	Uncens	1.00	1.00	1.09	1.02	1.00	1.00	1.26	1.06
	Cens 70	1.00	1.00	0.90	0.93	1.00	1.00	0.94	0.93
	Cens 90	1.00	1.00	1.00	1.01	1.00	1.00	1.01	1.04
$h=5$	Uncens	1.00	1.00	0.97	1.03	1.00	1.00	1.14	1.07
	Cens 70	1.00	1.00	1.12	1.08	1.00	1.00	1.14	1.14
	Cens 90	1.00	1.00	1.01	1.05	1.00	1.00	1.07	1.08
dof	ν								
$h=1$	Uncens	-	7.78	-	1000	-	1000	-	3.47
	Cens 70	-	1000	-	1000	-	1000	-	1000
	Cens 90	-	2.53	-	2.59	-		-	2.42
$h=3$	Uncens	-	4.45	-	4.48	-	3.35	-	3.41
	Cens 70	-	1000	-	3.82	-	4.18	-	4.10
	Cens 90	-	4.37	-	4.17	-	3.08	-	2.99
$h=5$	Uncens	-	3.36	-	3.34	-	2.80	-	2.79
	Cens 70	-	1000	-	1000	-	1000	-	1000
	Cens 90	-	2.99	-	2.70	-	2.70	-	2.85

Notes: Estimated parameters of the 100α=10% (Cens 90) and 100α=30% (Cens 70) and uncensored (Uncens) densities. Pre-pandemic sample (with dates referring to the outturn) is from 1967q1-2014q4 across h. Incl. pandemic estimates are simulated and involve appending a single error of -20% to the pre-pandemic sample of forecast errors. Second-release GDP estimates used to define the GDP growth outturns.

A.7 Two-sided variant of the censored LR test of Berkowitz

Here we set out a two-sided variant of the censored LR test proposed by Berkowitz (2001). This ignores the degree of forecast failure in both the left- and right-hand-side tails but importantly accounts for their frequency. For expositional ease, we suppress dependence of the PITs, z_t, on the forecast horizon, h.

Specifically, following Berkowitz (2001), take an inverse normal CDF transformation, Φ^{-1}, of the PITs to define $z_t^* = \Phi^{-1}(z_t)$; and define $z_{L,t}^* = \Phi^{-1}(z_{L,t})$ and $z_{U,t}^* = \Phi^{-1}(z_{U,t})$ such that:

$$z_{c,t}^* = z_t^* \text{ if } z_{L,t}^* \leq z_{t+h}^* \leq z_{U,t}^* \tag{A.14}$$

$$z_{c,t}^* = z_{L,t}^* \text{ if } z_t^* < z_{L,t}^* \tag{A.15}$$

$$z_{c,t}^* = z_{U,t}^* \text{ if } z_t^* > z_{U,t}^* \tag{A.16}$$

so that the log likelihood function for estimation of the mean and standard deviation, m and s, of z_t^*, which should be $(0, 1)$, respectively, under correct calibration, is given as:

$$\begin{aligned} L(m, s \mid z_{c,t}^*) = & \sum_{z_{L,t}^* < z_{c,t}^* < z_{U,t}^*} \log \frac{1}{s} \phi \left(\frac{z_{c,t}^* - m}{s} \right) \\ & + \sum_{z_{c,t}^* = z_{L,t}^*} \log \Phi \left(\frac{z_{L,t}^* - m}{s} \right) \\ & + \sum_{z_{c,t}^* = z_{U,t}^*} \log \left(1 - \Phi \left(\frac{z_{U,t}^* - m}{s} \right) \right). \end{aligned} \tag{A.17}$$

Therefore, a censored (or tail) LR test statistic can be constructed as:

$$LR_{tail} = -2(L(0, 1) - L(\widehat{m}, \widehat{s})), \tag{A.18}$$

that is distributed $\chi^2(2)$ under the null hypothesis that the censored density forecast is correctly calibrated (that is, $m = 0$ and $s = 1$). This two-degrees-of-freedom variant of Berkowitz's test (see Clements (2004)) does not test for independence in the PITs; we should not expect independence, under correct calibration, for forecast horizons greater than one.

A.8 Fitting uncensored and censored densities: Robustness

Here we report supplementary empirical results referred to in the main body of the paper.

Figure A3 shows the uncensored two-piece t and normal densities fitted to two-years-ahead UK forecast errors using the second-release GDP growth data as the outturn. Comparison with Figure A1, which shows analogous densities but with outturns measured using "mature" estimates of GDP, reveals that data revisions matter. Figure A3 indicates more skew to the forecast errors when second-release data are used as outturns. For the two-piece normal density, the skewness parameter diverges to 102 post-pandemic. For the two-piece t density, γ rises from 1.3 to 2.6.

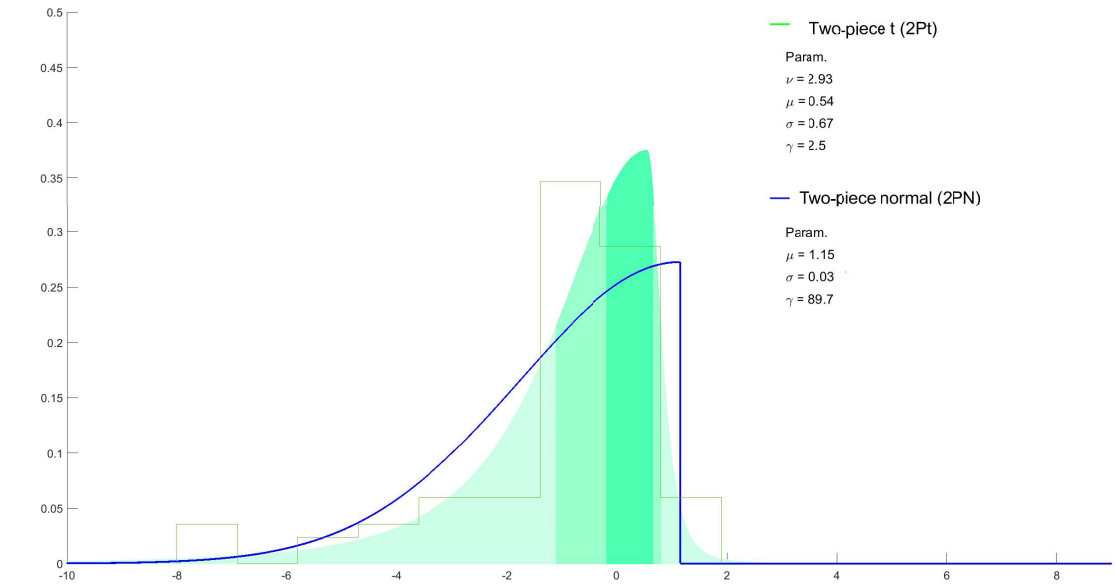
Figure A4 shows the censored two-piece t and normal densities fitted to forecast errors using the second-release GDP growth data as the outturn. Again we see much more skew than for the corresponding density estimates using final vintage data; cf. Figure 1 (in the main paper). Including pandemic data, the quadratic criterion, P_r, converges to a value of 0 but only for the two-piece t densities. For the two-piece normal densities, despite experimentation, it did not prove possible to obtain satisfactory estimation and $P_r > 0$ even as $r \to \infty$. The estimated densities failed to meet the requirements of a BCR (that is, the probability density of being at either censoring point should be equal); we therefore do not report them in Figure A4.

Figure A5 reports the censored densities fitted to UK GDP growth but using a 30 percent censoring region. This reveals yet more evidence for symmetry.

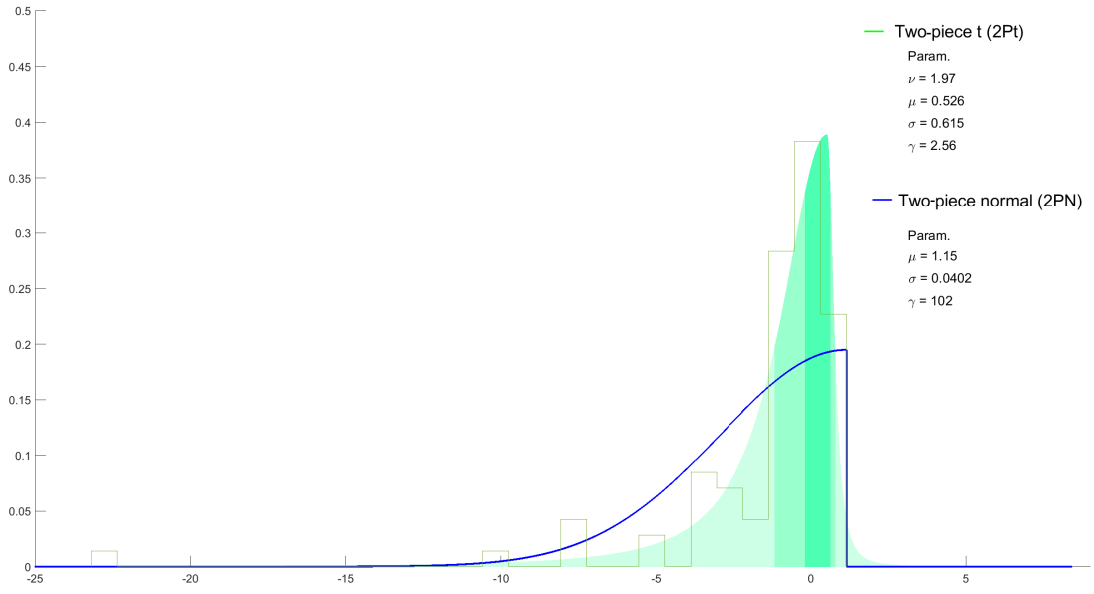
Figure A6 complements Figure 2 in the main paper, by exploring the properties of the censored densities fitted to US GDP growth errors when two simulated observations of -20 percent and +20 percent are added to the real FRB forecast error data. As anticipated, when the outliers fall in both tails, the uncensored density is more Gaussian and there is less evidence for skew than in Figure 2. Recall that the implication of the modest skew in Figure 2 is to allow the downside uncertainties to exceed the upside uncertainties.

Figure A7 illustrates the property, reported in the main paper: that the skew parameter of the censored 2PN often diverges when fitted recursively (out-of-sample) to the UK GDP errors. In contrast, the fat tails of the 2Pt prevent this from happening.

Figure A8, as referenced in the main paper, shows that the use of mature (specifically, 2018q4-vintage) GDP estimates rather than second-release GDP estimates to define the "outturn" has little effect on either the shape of the forecast error histograms or the fitted densities in the US; cf. Figures A2 and A8.



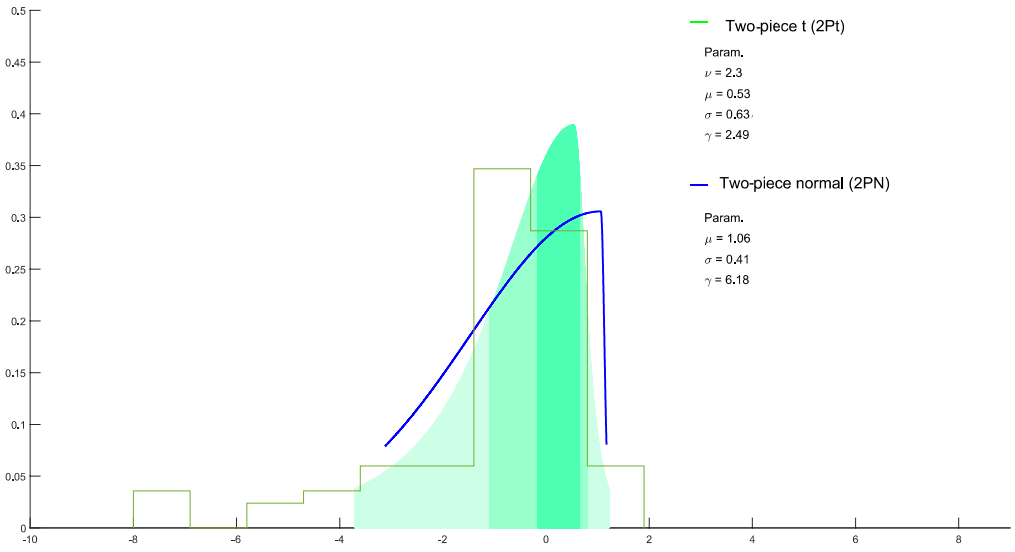
(a) Pre-pandemic



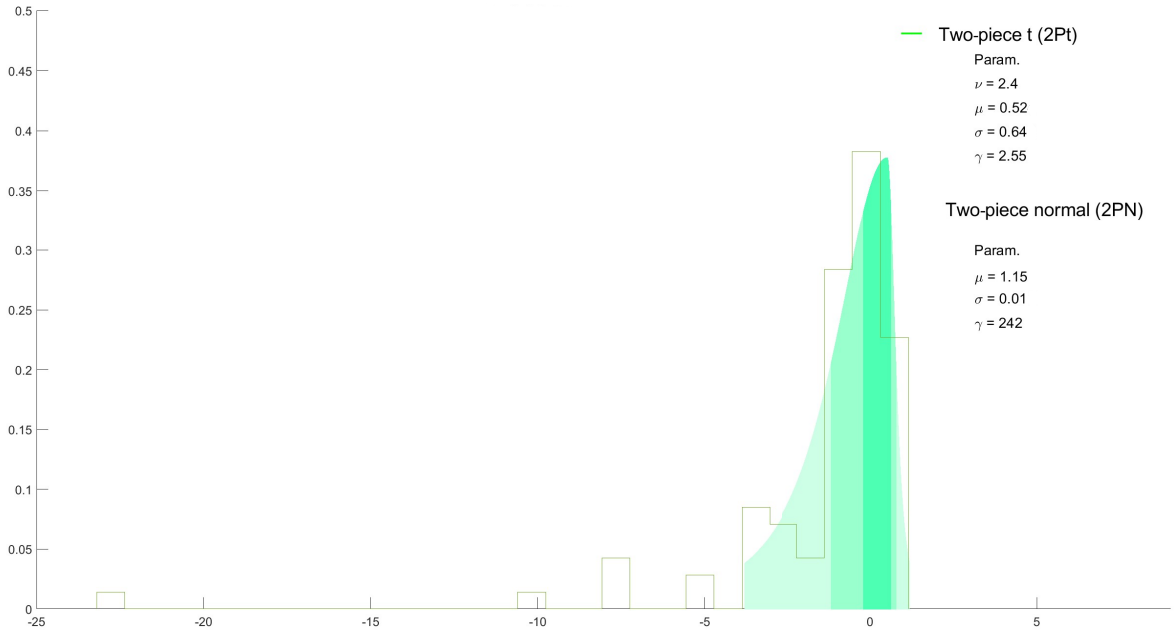
(b) Including pandemic

Figure A3: UK GDP Growth (two-years-ahead): MPC Forecast Error Histogram and Uncensored Two-Piece Normal and t Densities and Their Parameters

Note: Second-release GDP estimates used to define the "outturn." Pre-pandemic: 76 outturns used from 1999q4-2018q3. Including pandemic: 84 observations used from 1999q4-2020q3. The darkest shaded green region indicates the 30% best critical region of the 2Pt; the next band extends this to 60% with the remaining region of the density in the palest shade.



(a) Pre-pandemic

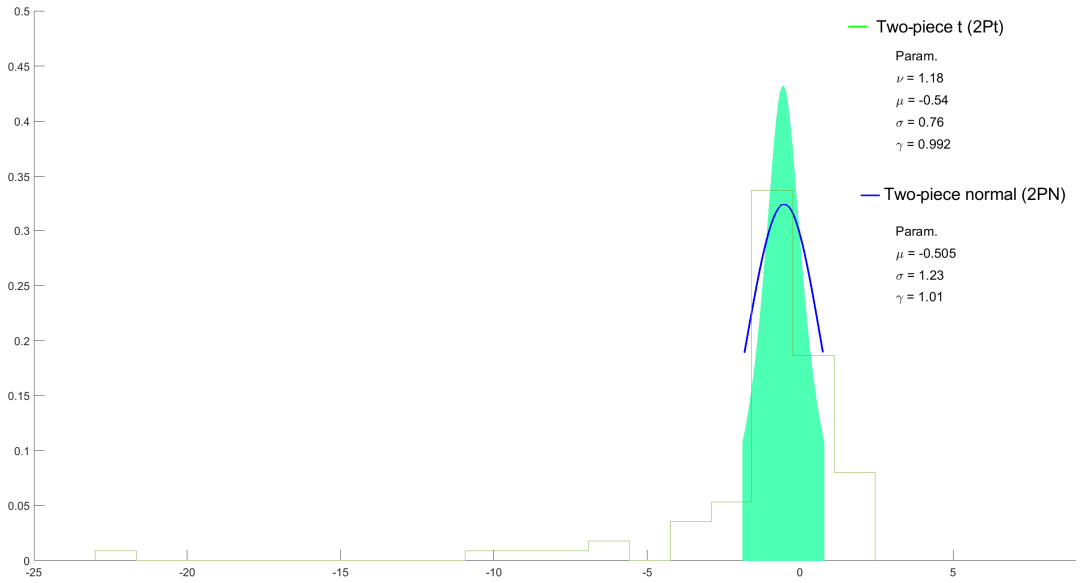


(b) Including pandemic

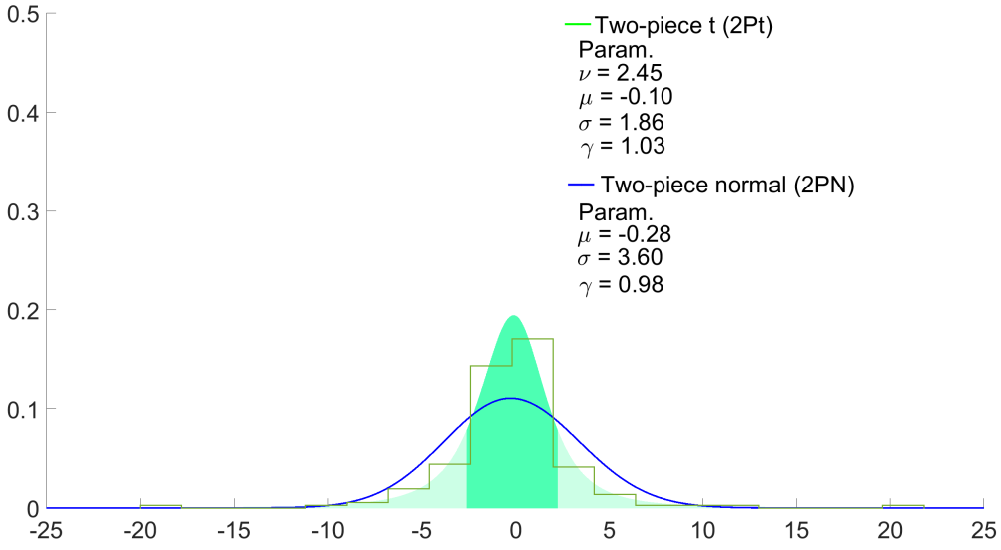
Figure A4: UK GDP Growth (two-years-ahead): Forecast Error Histogram and 10% Censored Two-Piece Normal and t Densities and Their Parameters Using L_A^C

Note: Second-release GDP estimates used to define the "outturn." Pre-pandemic: 76 outturns used from 1999q4-2018q3. Including pandemic: 84 observations used from 1999q4-2020q3. The darkest shaded green region indicates the 30% best critical region of the 2Pt; the next band extends this to 60% with the palest shade extending to 90%. Blue 2PN density not shown post-pandemic as the fixed-point algorithm did not converge.

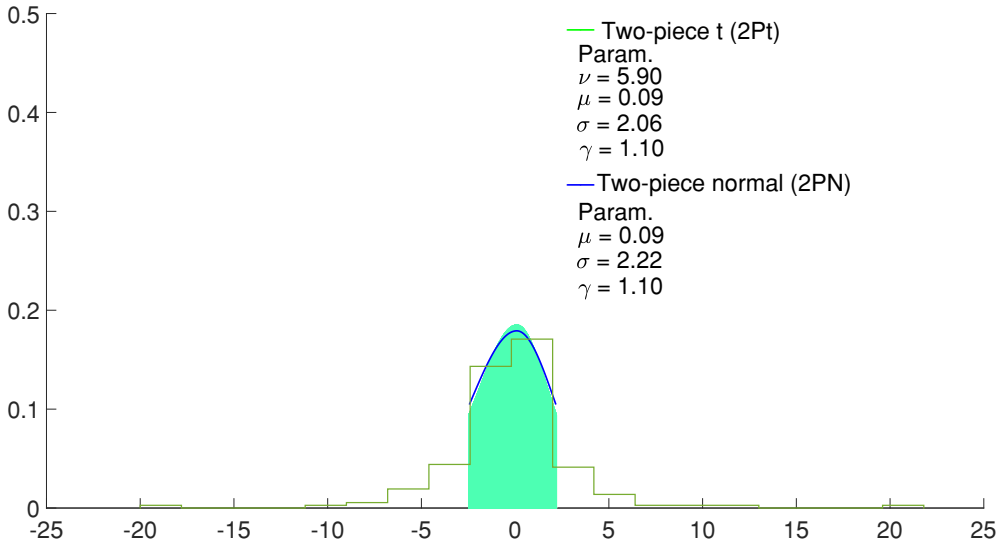
Figure A5: UK GDP Growth (two-years-ahead): Forecast Error Histogram and 30% Censored Two-Piece Normal and t Densities Using L_A^C Updated to 2020q3



Note: 84 including the pandemic forecast error observations used, 1999q4-2020q3. Mature GDP estimates used to define the "outturn." The shaded green region indicates the 70% best critical region of the 2Pt



(a) Uncensored



(b) 30% Censored

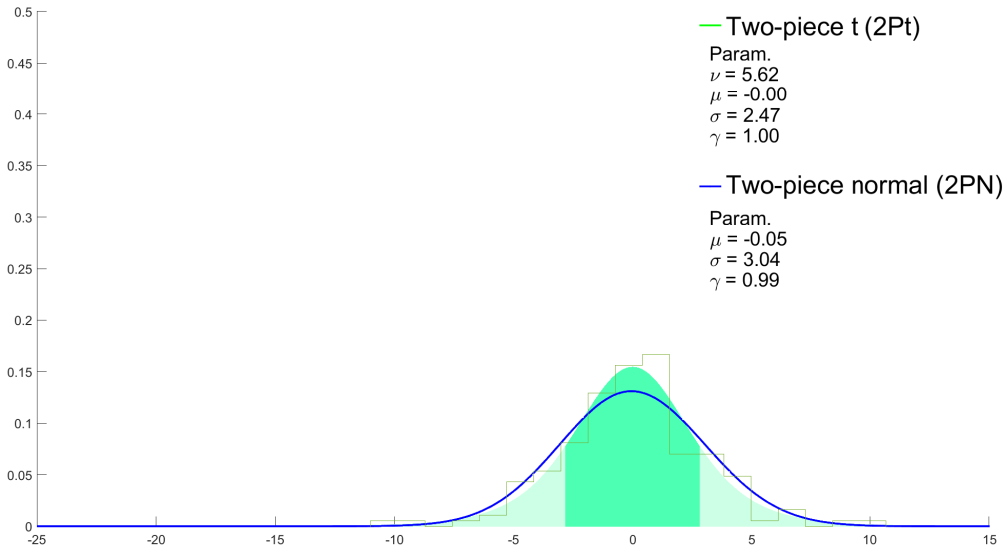
Figure A6: Simulated Post-pandemic US GDP Growth (one-year-ahead): FRB Forecast Error Histogram and Uncensored and 30% Censored Two-Piece Normal and t Densities Using L_A^C

Note: Second-release GDP estimates used to define the "outturn." Including pandemic sample from 1974q2-2014q4 plus two artificial observations of -20%. and +20%. The darkest shaded green region indicates the 70% best critical region of the 2Pt.

Figure A7: UK GDP Growth (two-years-ahead): Real-time Estimates of the Standard Deviation and Skew Parameters Using the 10% Censored 2PN Density

Note: Second-release data used to define GDP outturns. 84 including the pandemic forecast error observations used, 1999q4-2020q3. Dates refer to the target, with the forecast made two years previously.

Figure A8: US GDP growth (one-year-ahead): FRB Forecast Error Histogram and Uncensored Two-Piece Normal and t Densities Using Mature Outturns

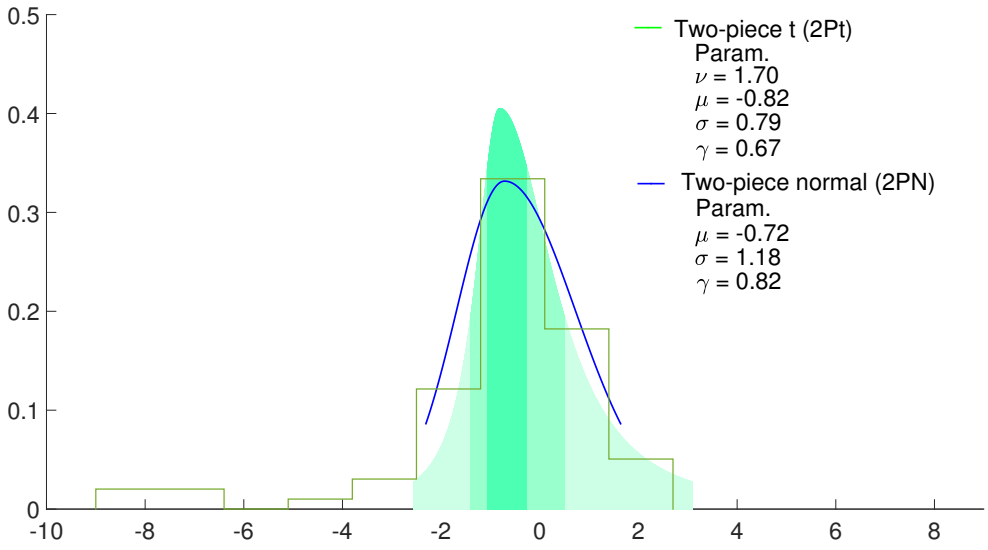


Note: 2018q4-vintage GDP estimates used to define the mature outturns for GDP growth. Forecast error sample is 1974q2-2014q4 (pre-pandemic). The darkest shaded green region indicates the 70% best critical region of the 2Pt.

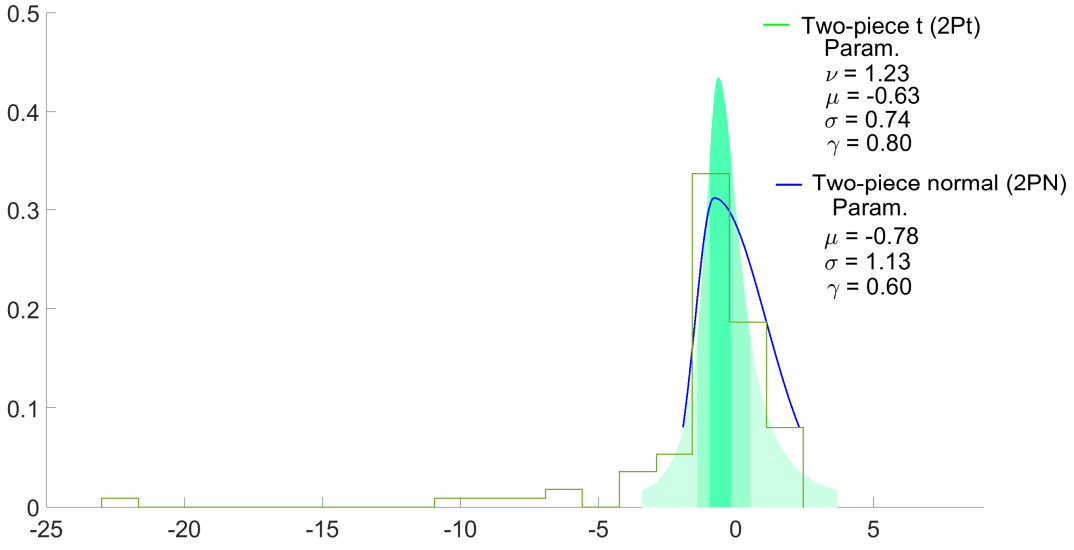
A.8.1 Censored densities: Use of L_B^C

Figure A9 shows the consequences of fitting L_B^C rather than L_A^C to the UK (two-years-ahead) GDP forecast errors, with mature data used to define GDP outturns. The quadratic criterion, P_r, converges to a value of 0. Recall that L_A^C distinguishes the lower from the upper tail, with each having its own probability. This means that the process of fitting is likely to place some observations in each tail rather than locating all the censored observations in only one of the tails as in L_B^C - our focus here. The expected number of observations in each tail depends on the skew parameter.

Figure A9 indicates both less evidence for asymmetry relative to L_A^C (seen in Figure 1) and that the nature of the observed asymmetry for the 2Pt has switched from right to left skew. This is because all of the censored observations associated with the global financial crisis and the pandemic are now in the left tail. As a result, the distribution is more symmetric because no attempt is made to place any censored observations in the right-hand tail, as in Figure 1 using L_A^C. We also find, however, that even when no effort is made to accommodate the recession (given that the recessionary data are censored), a low number of degrees of freedom is estimated.



(a) Pre-pandemic



(b) Including pandemic

Figure A9: UK GDP Growth (two-years-ahead): MPC Forecast Error Histogram and 10% Censored Two-Piece Normal and t Densities and Their Parameters Using L_B^C

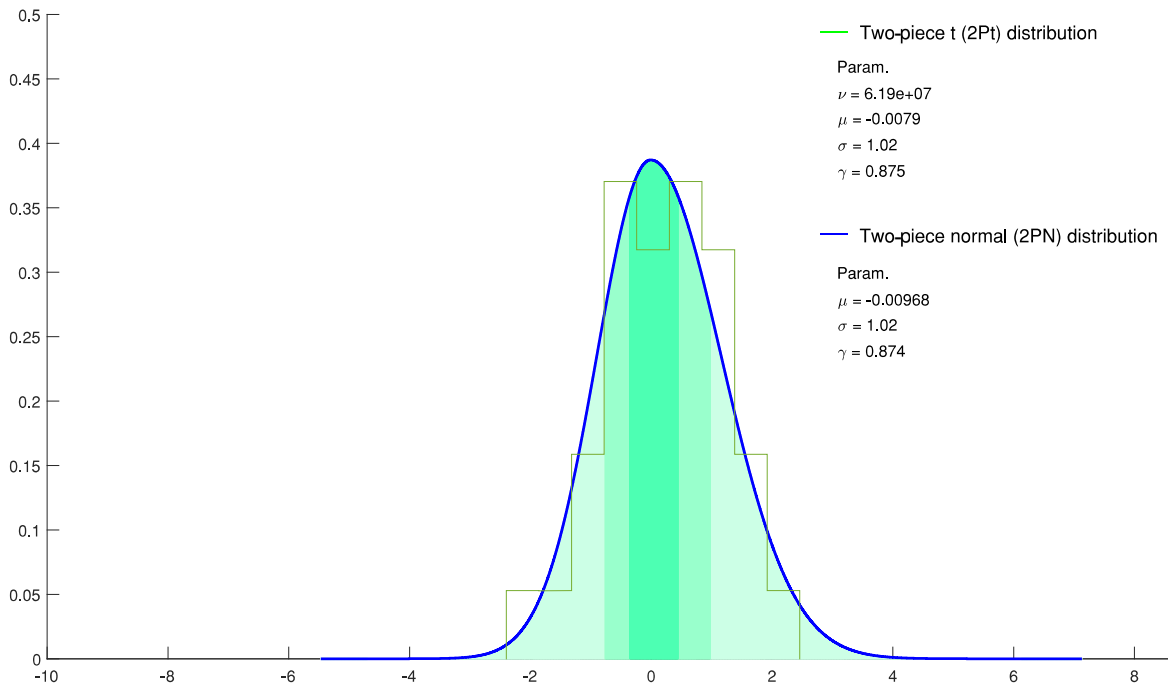
Note: Mature GDP estimates used to define the "outturn." Pre-pandemic: 76 outturns used from 1999q4-2018q3. Including pandemic: 84 observations used from 1999q4-2020q3. The darkest shaded green region indicates the 30% best critical region of the 2Pt; the next band extends this to 60% with the palest shade extending to 90%.

A.8.2 Densities using specific windows of data

Figures A10-A11 follow in the spirit of practice at the Bank of England by plotting some illustrative (uncensored) densities fitted to specific (rolling) samples of UK GDP growth forecast error data.

We select the sample period carefully/subjectively, aware of the effects of the global financial crisis in 2008 on the GDP forecast errors. Accordingly, Figure A10 considers a sample of MPC (two-years-ahead) forecast error data before the global financial crisis, while Figure A11 considers a sample after the global financial crisis but before the pandemic. Experimentation revealed that the choice of estimation window for these uncensored densities could have a large effect on the shape of the densities fitted to the GDP forecast errors. The dates referenced in the figures refer to outturns, with the forecasts made two years previously.

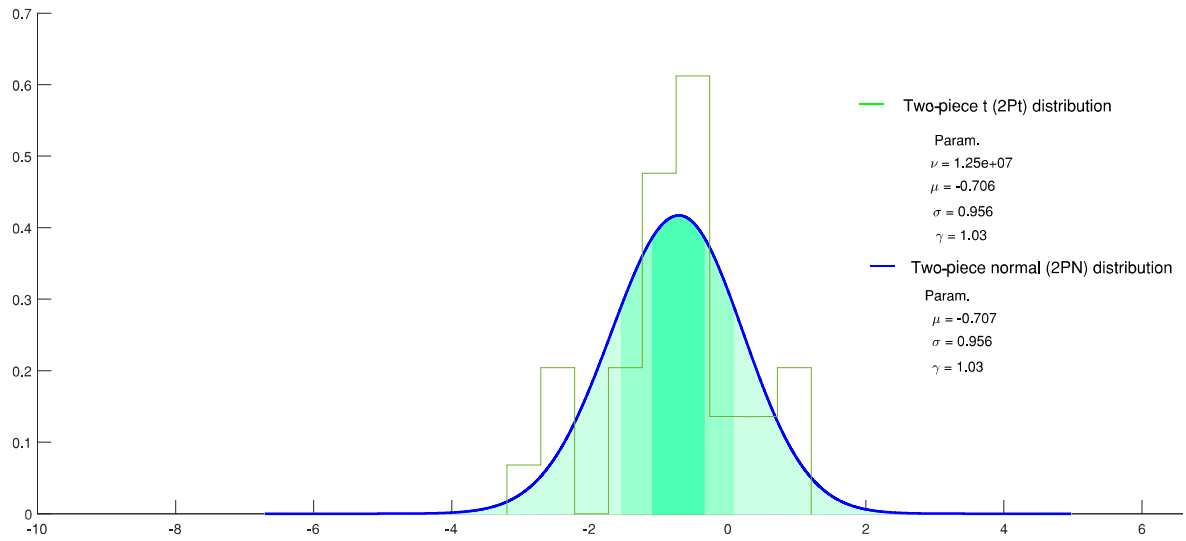
Figure A10: UK GDP Growth (two-years-ahead): Forecast Error Histogram and Uncensored Two-Piece Normal and t Densities and Their Parameters Fitted to Error Data from 1999q4-2008q2



A.9 Tracking the temporal evolution of the out-of-sample censored intervals

Here we take the out-of-sample censored density forecasts computed in Section 8 and look at their evolution over time. In so doing, we shed light on the temporal behavior of *ex ante* expectations of *knowable* uncertainty, defined as the censoring bounds $y_{U,t} - y_{L,t}$ at 10 percent. For parsimony, Figure A12 plots these 10 percent censoring bounds or 90 percent forecast intervals for just the more flexible censored 2Pt density, as computed in real time from the historical forecast errors for UK GDP growth. This censored density is compared with censoring bounds extracted from the rolling (uncensored) Gaussian density (rolling N) and the censoring bounds extracted from the MPC's own judgment-informed 2PN density forecasts, as published in the *Inflation* now *Monetary Policy Reports*. Figure A13 provides an analogous plot of the 70 percent forecast intervals for US GDP growth.

Figure A11: UK GDP Growth (two-years-ahead): Forecast Error Histogram and Uncensored Two-Piece Normal and t Densities and Their Parameters Fitted to Error Data from 2011q2-2018q3



Alongside these *ex ante* censoring intervals, we plot the *ex post* forecast error - the outturn.

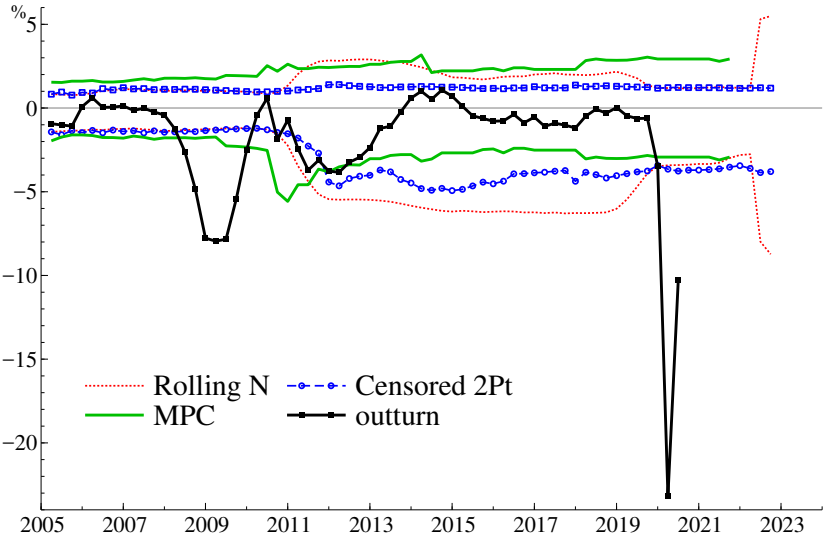
Turning to the UK first, Figure A12 reveals that the 90 percent intervals from all three approaches widened appreciably in the aftermath of the negative forecast errors observed over the period of the global financial crisis. The MPC judgment-based forecasts picked up this increased uncertainty earlier than the data-based estimators, which, as expected, adjust only with a lag. For MPC and censored 2Pt, this increase in *ex ante* uncertainty is largely explained by a fall in the lower bound. This fits a narrative of heightened downside risk after the global financial crisis. This is confirmed by Figure A14, which plots the recursively estimated skew parameters. Extending the production of the error intervals beyond periods of time when outturn data are available, Figure A12 reveals that the censoring bounds from the uncensored Gaussian density widen sharply, once the large negative outturns observed in 2020 due to the pandemic are considered. In contrast, the censored 2Pt interprets the extreme outturns of the COVID-19 period as outliers. Accordingly, it places them in the censored region of the density so that similar estimates are observed pre-pandemic (also see Figure A14). 2Pt is temporally more stable.

For the US, in Figure A13 we again observe higher uncertainty when an uncensored

Gaussian density is fitted to a rolling 20-year window of forecast errors. As in the UK, we see the width of the 70 percent interval from rolling N widen in the aftermath of the negative forecast errors made during the global financial crisis. We also see wider intervals than the censored 2Pt in the early sample from 1995 to 2000. By contrast, the censoring intervals from the censored 2Pt density are both narrower and again more stable over time. As noted by Reifschneider and Tulip (2019), forecast-error-based estimates of uncertainty are sensitive to the sample period. But our results show that this sensitivity and the volatility of the estimates diminish when a censored density is fitted to the error data. Figure A15 also shows how the standard deviation of the 2PN density is very stable over time, in contrast to the temporal changes seen in the uncensored rolling N density. It is this temporal variation that motivates the use of models of time variation in forecast error variances, as in, for example, Clark, McCracken, and Mertens (2020).

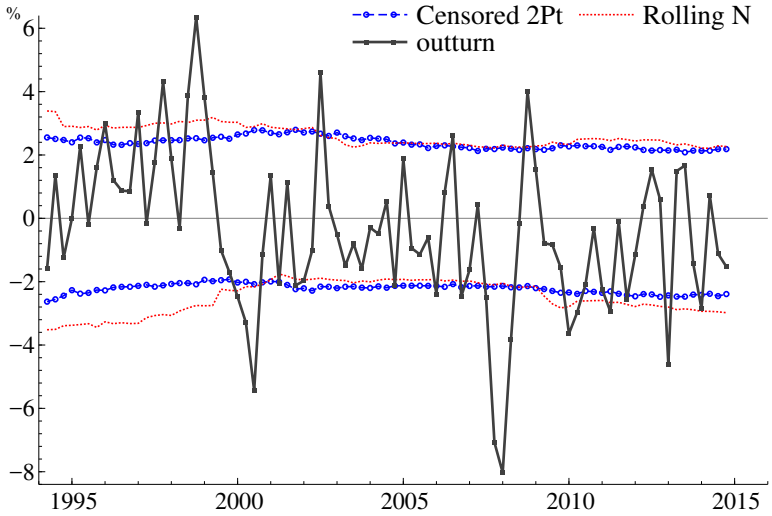
Our results therefore suggest that censored densities exhibit more temporal stability in their moments, notably the variance, than when an uncensored Gaussian density is fitted unconditionally to a rolling window of forecast errors. This is consistent with arguments in Orlik and Veldkamp (2014): real-time estimation of densities with non-normal tails is prone to large changes in the variance - new observations "wag the tail" of the whole density. Since variance is the expected squared distance from the mean, changes in the probabilities of outliers have large effects on the conditional variance. There is less need to model time variation in forecast error standard deviations or variances when outliers are censored.

Figure A12: Real-time UK GDP 90% *Ex Ante* Forecast Error Intervals, $y_{U,t} - y_{L,t}$, and *Ex Post* Forecast Errors



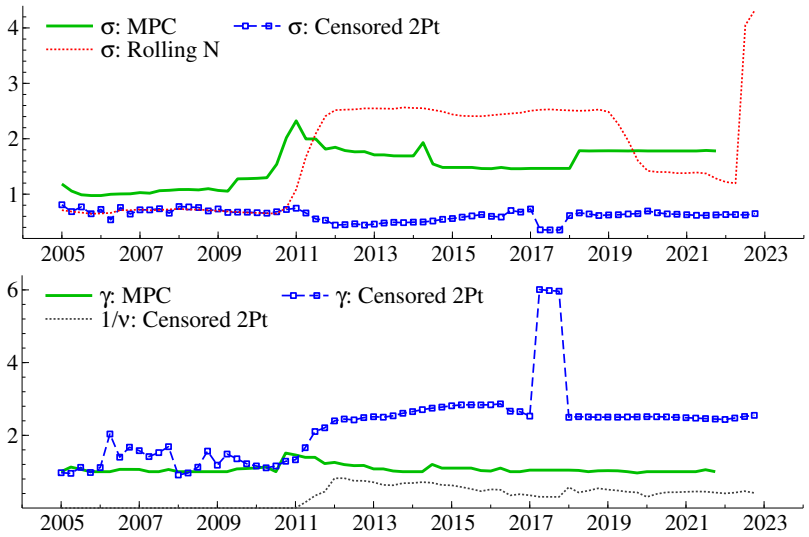
Notes: Two-years-ahead forecast errors for outturns from 2005q1-2020q4; second-release GDP estimates used to define outturns. Dates refer to the *ex post* forecast error outturn, with the forecast made two years previously.

Figure A13: Real-time US GDP 70% *Ex Ante* Forecast Error Intervals, $y_{U,t} - y_{L,t}$, and *Ex Post* Forecast Errors



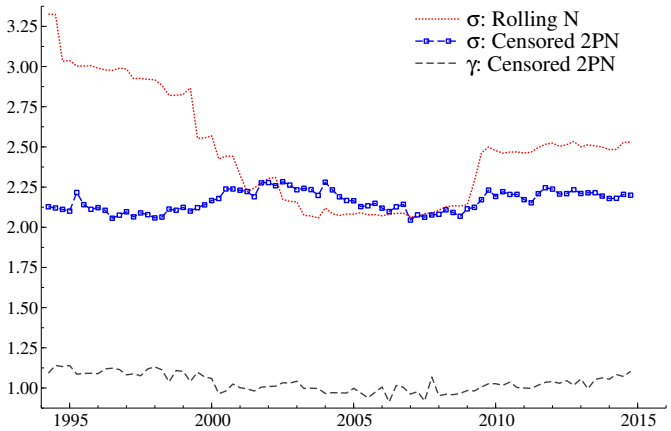
Notes: One-year-ahead forecast errors for outturns from 1994q2-2014q4; second-release GDP estimates used to define outturns. Dates refer to the *ex post* forecast error outturn, with the forecast made one year previously.

Figure A14: Real-time Data-Based and MPC Parameters for the UK GDP Error Density



Notes: Two-years-ahead forecasts. Second-release GDP estimates used to define outturns. Dates refer to the outturn, with the forecast made two years previously. The censored density is censored at α=0.1.

Figure A15: Real-time Data-Based Parameters for the US GDP Error Density



Notes: One-year-ahead forecasts. Second-release GDP estimates used to define outturns. Dates refer to the outturn, with the forecast made one year previously. The censored densities are censored at α=0.3.

A.10 Appendix References

Adrian, Tobias, Nina Boyarchenko, and Domenico Giannone (2019). "Vulnerable growth." *American Economic Review*, 109(4), pp. 1263-1289. doi:10.1257/aer.20161923.

Arellano-Valle, Reinaldo B., Héctor W. Gómez, and Fernando A. Quintana (2005). "Statistical inference for a general class of asymmetric distributions." *Journal of Statistical Planning and Inference*, 128, pp. 427-443. doi:10.1016/j.jspi.2003.11.014.

Azzalini, Adelchi (1985). "A class of distributions which includes the normal ones." *Scandinavian Journal of Statistics*, 12(2), pp. 171-178. URL https://www.jstor.org/stable/4615982.

Azzalini, Adelchi (2018). "Package 'sn' - The R Project for Statistical Computing.". URL https://cran.r-project.org/web/packages/sn/sn.pdf.

Azzalini, Adelchi and Reinaldo B. Arellano-Valle (2013). "Maximum penalized likelihood estimation for skew-normal and skew-t distributions." *Journal of Statistical Planning and Inference*, 143, pp. 419-433. doi:10.1016/j.jspi.2012.06.022.

Azzalini, Adelchi and Antonella Capitanio (2003). "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution." *Journal of the Royal Statistical Society: Series B*, 65, pp. 367-389. doi:10.1111/1467-9868.00391.

Berkowitz, Jeremy (2001). "Testing density forecasts with applications to risk management." *Journal of Business and Economic Statistics*, 19, pp. 465-474. doi:10.1198/07350010152596718.

Clark, Todd E., Michael W. McCracken, and Elmar Mertens (2020). "Modeling time-varying uncertainty of multiple-horizon forecast errors." *Review of Economics and Statistics*, 102(1), pp. 17-33. doi:10.1162/rest_a_00809.

Clements, Michael P. (2004). "Evaluating the Bank of England density forecasts of inflation." *Economic Journal*, 114(498), pp. 844-866. doi:10.1111/j.1468-0297.2004.00246.x.

Elder, Rob, George Kapetanios, Tim Taylor, and Tony Yates (2005). "Assessing the MPC's Fan Charts." *Bank of England Quarterly Bulletin*, 45, pp. 326-348. URL https://www.bankofengland.co.uk/quarterly-bulletin/2005/q3/assessing-the-mpcs-fan-charts

Fernandez, Carmen and Mark F.J. Steel (1998). "On Bayesian modelling of fat tails and skewness." *Journal of the American Statistical Association*, 93, pp. 359-371.

doi:10.1080/01621459.1998.10474117.

Jones, M. Chris and Arthur Pewsey (2009). "Sinh-arcsinh distributions." *Biometrika*, 96, pp. 761-780. doi:10.1093/biomet/asp053.

Mudholkar, Govind S. and Alan D. Hutson (2000). "The epsilon-skew-normal distribution for analyzing near-normal data." *Journal of Statistical Planning and Inference*, 83, pp. 291-309. doi:10.1016/S0378-3758(99)00096-8.

Pesaran, M. Hashem, Andreas Pick, and Mikhail Pranovich (2013). "Optimal forecasts in the presence of structural breaks." *Journal of Econometrics*, 177, pp. 134-152. doi:10.1016/j.jeconom.2013.04.002.

Ramirez-Cobo, Pepa, Rosa E. Lillo, Simon Wilson, and Michael P. Wiper (2010). "Bayesian inference for double Pareto lognormal queues." *Annals of Applied Statistics*, 4(3), pp. 1533-1557. doi:10.1214/10-AOAS336.

Reifschneider, David L. and Peter Tulip (2019). "Gauging the uncertainty of the economic outlook using historical forecasting errors: The Federal Reserve's approach." *International Journal of Forecasting*, 35(4), pp. 1564-1582. doi:10.1016/j.ijforecast.2018.07.016.

Rubio, Francisco J. and Mark F.J. Steel (2014). "Inference in two-piece location-scale models with Jeffreys priors, with discussion." *Bayesian Analysis*, 9, pp. 1-22. doi:10.1214/13-BA849.

Rubio, Francisco J. and Mark F.J. Steel (2015). "Bayesian modelling of skewness and kurtosis with two-piece scale and shape distributions." *Electronic Journal of Statistics*, 9, pp. 1884-1912. doi:10.1214/15-EJS1060.

Sartori, Nicola (2006). "Bias prevention of maximum likelihood estimates for scalar skew normal and skew t distributions." *Journal of Statistical Planning and Inference*, 136(12), pp. 4259-4275. doi:10.1016/j.jspi.2005.08.043.

Wallis, Kenneth, F. (2014). "The two-piece normal, binormal, or double Gaussian distribution: its origin and rediscoveries." *Statistical Science*, 29(1), pp. 106-112. https://doi.org/10.1214/13-STS417.

Zhu, Dongming and John W. Galbraith (2010). "A generalized asymmetric Student-t distribution with application to financial econometrics." *Journal of Econometrics*, 157, pp. 297-305. doi:10.1016/j.jeconom.2010.01.013.