Norges Bank

14 | 2024

Working Paper

Quantile combination: An application to US GDP growth forecasts

Norges Bank Research

Authors:
Knut Are Aastveit
Saskia ter Ellen
Giulia Mantoan

Keywords
Density forecasts, forecast combinations, quantile regression, downside risk

Working papers fra Norges Bank, fra 1992/1 til 2009/2 kan bestilles på e-post: servicesenter@norges-bank.no

Fra 1999 og senere er publikasjonene tilgjengelige på www.norges-bank.no

Working papers inneholder forskningsarbeider og utredninger som vanligvis ikke har fått sin endelige form. Hensikten er blant annet at forfatteren kan motta kommentarer fra kolleger og andre interesserte. Synspunkter og konklusjoner i arbeidene står for forfatternes regning.

Working papers from Norges Bank, from 1992/1 to 2009/2 can be ordered by e-mail: servicesenter@norges-bank.no

Working papers from 1999 onwards are available on www.norges-bank.no

Norges Bank's working papers present research projects and reports (not usually in their final form) and are intended inter alia to enable the author to benefit from the comments of colleagues and other interested parties. Views and conclusions expressed in working papers are the responsibility of the authors alone.

ISSN 1502-8143 (online)

ISBN 978-82-8379-334-5 (online)

Quantile combination: An application to US GDP growth forecasts*

Knut Are Aastveit† Saskia ter Ellen‡ Giulia Mantoan§

July 3, 2024

Abstract

We propose an easy-to-implement framework for combining quantile forecasts, applied to forecasting GDP growth. Using quantile regressions, our combination scheme assigns weights to individual forecasts from different indicators based on quantile scores. Previous studies suggest distributional variation in forecasting performance of leading indicators: some indicators predict the mean well, while others excel at predicting the tails. Our approach leverages this by assigning different combination weights to various quantiles of the predictive distribution. In an empirical application to forecast US GDP growth using common predictors, forecasts from our quantile combination outperform those from commonly used combination approaches, especially for the tails.

Keywords: *Density forecasts; forecast combinations; quantile regressions; downside risk*

JEL classification: *C32, C53, E37*

*This paper should not be reported as representing the views of Norges Bank or Bank of England. The views expressed are those of the authors and do not necessarily reflect those of Norges Bank or Bank of England. The authors would like to thank Laurent Ferrara, Gergerly Gánics, Malte Knüppel, Michael McCracken, James Mitchell, Tatevik Sekhposyan, Herman van Dijk, an anonymous referee for the Norges Bank Working Paper series, and seminar and conference participants at Norges Bank, Bank of England, the 2020 Joint Statistical Meeting, the 2020 SNDE symposium, the 2020 International Symposium on Forecasting, the 2021 IAAE conference, the Fall 2021 Federal Reserve Bank of St. Louis Applied Time Series Econometrics Workshop, IIF MacroFor, the 2022 Conference on real-time data analysis, methods and applications at the Federal Reserve Bank of Cleveland and the 2023 ECB Conference on Forecasting Techniques. This paper is part of the research activities at the Centre for Applied Macroeconomics and Commodity Prices (CAMP) at the BI Norwegian Business School. This work was supported by the Economic and Social Research Council DTC grant number ES/J500203/1.

†Norges Bank & Centre for Applied Macroeconomics and Commodity Prices (CAMP), BI Norwegian Business School; Knut-Are.Aastveit@Norges-Bank.no

‡Vrije Universiteit Amsterdam; S.ter.Ellen@vu.nl

§Bank of England; Giulia.Mantoan@bankofengland.co.uk

1 Introduction

Uncertainty and downside risk play a prominent role in economic forecasting and policy making. To reflect uncertainty around mean predictions, it has become common practice for economic forecasters, particularly in central banks, to provide density forecasts. In recent years, policymakers became particularly interested in quantifying macroeconomic downside tail risk.[1] The most prominent contribution to this aim comes from Adrian et al. (2019), who introduce the concept of GDP-at-risk as a macroeconomic counterpart to the financial value-at-risk. Adrian et al. (2019) study the distribution of macroeconomic risk by estimating a quantile forecast regression of real GDP growth over the next year for various quantiles and show that financial conditions – captured by the National Financial Conditions Index (NFCI) – are particularly informative about macroeconomic downside risk.[2]

Although it is practical to rely on a single indicator to predict macroeconomic downside risk, this approach has disadvantages. The complexity of the economy might not be captured by a single indicator, and neglecting other predictors could pose a risk in itself. A vast amount of research has shown that a variety of macroeconomic and financial variables contain predictive information about future economic recessions and downturns, see e.g. Marcellino (2006) and Liu and Moench (2016) for an overview. Reichlin et al. (2020) find that financial conditions offer limited additional information on economic downturns beyond what is already captured by real economic indicators. This finding is further supported by Amburgey and McCracken (2023a), who document that growth-at-risk is essentially investment-at-risk.[3]

To leverage the predictive power of multiple models and predictors, one can turn to forecast density combination methods. However, these approaches typically assign a single weight to the entire predictive distribution for each model, not capturing the findings in Manzan (2015) that some models may be good at forecasting the mean of the distribution but perform poorly in the tails, or vice versa.

We therefore propose a coherent methodology to construct density forecasts that incorporates the heterogeneity in accuracy across regions of the forecast distribution from multiple sources. The *quantile combination* approach we propose can achieve more accurate density forecasts by assigning weights to individual forecasts from different indicators

[1] In the US, the Federal Open Market Committee (FOMC) commonly discusses downside risks to growth in FOMC statements, with the relative prominence of this discussion fluctuating with the business cycle. More generally, macroeconomic downside risk has also been the focus of recent publications and speeches by policy institutions such as the International Monetary Fund (IMF), Bank of Canada and Bank of England.

[2] This has led to a surge of interest in growth-at-risk (e.g. Coe and Vahey, 2020; Reichlin et al., 2020; Carriero et al., 2022; Clark et al., 2023; Brownlees and Souza, 2021; Amburgey and McCracken, 2023b).

[3] Amburgey and McCracken (2023a) argue that if financial conditions indicate that US real GDP growth will fall into the lower tail of its conditional distribution, the main contributor is a decline in investment.

based on quantile scores.

First, we generate individual forecasts using quantile regression models. Next, we combine these individual forecasts using a novel quantile combination approach, where each quantile of the combined density forecast is constructed as a weighted combination of the individual forecasts for the corresponding quantile. To accommodate the heterogeneity in forecast accuracy across different models and parts of the distribution, we allocate quantile-specific weights from each model using the quantile score introduced by Gneiting and Ranjan (2011).[4] In a final step, after obtaining the combined quantiles of the predictive distribution, we follow Adrian et al. (2019) and fit the skew t-distribution developed by Azzalini and Capitanio (2003) to recover the full probability density function.

We demonstrate the usefulness of this quantile combination approach by forecasting the real GDP growth rate for the Unites States for the period 1993Q1-2019Q4 using a real-time dataset. We combine forecasts from $K = 9$ quantile regression models. Each quantile regression model includes lagged GDP growth and one additional predictor (with lags). Motivated by Adrian et al. (2019) and the extensive literature on predicting economic recessions, we include the NFCI, the University of Michigan Consumer Sentiment Index (ICS), unemployment rate (U), a credit spread (CrSpread) that measures the difference between BAA corporate bond yield and the 10 year treasury yield, residential investment (ResInv), new housing permits (Permit), total (PCE) and durable personal consumption expenditures (PCEDG) and industrial production (INDPRO). The results reveal a substantial heterogeneity in the predictive performance over the various quantiles and forecast horizons. None of the individual models perform equally well in forecasting over all the quantiles and for both forecasting horizons. By providing a flexible way to account for this heterogeneity, our method leads to density forecast that are more accurate than forecasts using a single predictor or generated by traditional combination methods.

Our quantile combination approach consistently provides well-calibrated densities and outperforms other combination approaches. Our findings are robust across various specifications, and remain robust when extending the evaluation sample to include the COVID-19 pandemic period, or when estimating models using a rolling window instead of an expanding window. This suggests that our results are not sensitive to changes in the sample period or the estimation method, reinforcing the reliability and stability of our findings.

Moreover, the forecasting performance of our quantile combination approach is robust to the choice of benchmark vintage for the "true" measure of GDP, something that can significantly impact empirical results, as highlighted in studies such as Croushore and Stark (2001) and Stark and Croushore (2002). As data revisions have a greater impact on the tail of the distribution than on the center, inference on quantiles could be particularly

[4]The quantile score is a strictly proper scoring rule that represents a weighted version, or decomposition, of the continuous ranked probability score (CRPS).

sensitive to the choice of vintage compared to standard point or density forecasts. Indeed, we show that the best-performing predictor for individual forecasting models varies across different benchmark vintages. The flexibility of our approach allows weights to adjust based on forecasting performance for different parts of the distribution, depending on the specific benchmark vintage selected.

We contribute to the findings of earlier forecast combination and GDP-at-risk literature in several ways.

First, we show that forecasts from our quantile combination approach outperform forecasts from commonly used combination approaches, including Bayesian Model Averaging (BMA), optimal combination of density forecasts (OptComb) as suggested by Hall and Mitchell (2007) and Geweke and Amisano (2011), and equal weights (EQ). This holds irrespective of using the CRPS or a quantile-weighted version of the CRPS that emphasizes performance in the center, left, or right tail of the distribution as a measure of forecast accuracy. Importantly, the relative gains in forecasting performance from our model are not specific to particular regions of the distribution or limited to specific subperiods within our forecasting sample. Instead, we find a steady improvement over time across all quantiles of the GDP distribution.

Second, we complement the findings from Adrian et al. (2019) by showing that besides the NFCI, other variables are also informative about future downside macroeconomic risk. Specifically, we find that incorporating variables such as residential investments, building permits and a consumer sentiment index leads to more accurate forecasts for the lower quantiles of the GDP distribution compared to quantile regressions that include only the NFCI.

Finally, our paper is also related to Opschoor et al. (2017), who assess the merits of density forecast combination schemes that assign weights to individual density forecasts based on the censored likelihood scoring rule of Diks et al. (2011) and the CRPS of Gneiting and Ranjan (2011). They apply this approach to measure downside risk in equity markets using individual volatility models. Besides focusing on GDP growth rather than on equity markets, our paper differs from theirs in two important aspects: we assign weights to individual forecasts based on quantile scores, and we do not solely focus on the lower tail of the distribution but aim to obtain density forecasts that are more accurate for all parts of the distribution.

The rest of the paper is organized as follows: Section 2 presents our quantile combination approach and the individual quantile regression models. Section 3 presents the data set we use and results from our empirical application. Section 4 concludes.

2 Econometric framework

In this section we describe our quantile forecast combination approach. Our combination approach aims to obtain overall more accurate density forecasts, by assigning a set of combination weights to the various quantiles of the individual forecasts. To achieve this goal, we first produce individual forecasts using quantile regression models, outlined in section (2.1). Then we combine the various individual forecasts using a novel quantile combination approach, where each quantile of the combined density forecast is constructed as a weighted combination of the individual forecasts for the corresponding quantile, detailed in section (2.2). Then, we build the predictive distributions fitting a skew t-distribution to our combined quantiles, discussed in (2.3). Finally we describe alternative combination approaches used in the literature, to which we will ultimately compare the out-of-sample performance of our proposed quantile combination approach.

2.1 Quantile regression models

Quantile regression, popularized in economics by Koenker and Bassett Jr (1978), generalizes the traditional least squares regression by fitting a distinct regression line for each quantile of the distribution of the variable of interest. In principle, we would like to know the entire conditional distribution function that relates the dependent variable with the predictors. Quantile regression approximates this by minimizing sums of asymmetrically weighted absolute residuals.

In this paper we forecast GDP growth with an autoregressive distributed lag (ARDL) model:

$$y_{t+h,q,k} = \boldsymbol{x}'_{t,k}\boldsymbol{\beta}_q + \varepsilon_{t+h} \tag{1}$$

where $\boldsymbol{x}'_{t,k}$ is the vector of lagged values of y_t, for $t = 1, \ldots, T$ (with maximum lag r) and of one of the $k = 1, \ldots, K$ predictors (with maximum lag p). In our empirical application, the number of lags p and r are selected using the BIC selection criterion with a maximum of four lags. Traditionally, quantile regression estimation for $\boldsymbol{\beta}_q$ proceeds by minimizing the quantile weighted absolute value of errors:

$$\hat{\boldsymbol{\beta}}_q = \min_{\beta_q} \sum_{t=1}^{T-h} \left(q \cdot I_{y_{t+h} \geq x_t\beta} |y_{t+h} - \boldsymbol{x}'_{t,k}\boldsymbol{\beta}_q| + (1-q) I_{y_{t+h} < x_t\beta} |y_{t+h} - \boldsymbol{x}'_{t,k}\boldsymbol{\beta}_q| \right) \tag{2}$$

and $I(\cdot)$ denotes the usual indicator function.[5] The set of quantiles provides a more complete description of the response distribution than the mean, making the quantile regression an important alternative to classical mean regression. Moreover, $q = 1, \ldots, 17$ denotes the respective quantile, set to 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65,

[5] To avoid quantile-crossing, we monotonically rearrange the quantile forecasts $\hat{y}_{t+h,q,k}$ following Chernozhukov et al. (2010).

70, 75, 80, 85, 90 percentiles in our empirical application. We decide to ignore more "extreme" quantiles (below 10 and above 90) for two reasons: first, it is well known that estimation error increases for very rare events; second, this estimation error will impact the combination technique we propose since it is based on the accuracy at these quantiles (corresponding to one or two events in the time series). We will focus on forecast horizons $h = \{1, 4\}$ in our empirical application.

In the rest of the paper, we will denote $\hat{y}_{t+h,q,k}$ as the quantile forecasts, $f(y)_{t+h,q,k}$ as its predictive probability density function (pdf) and its cumulative counterpart as $F(y)_{t+h,q,k}$.

2.2 Quantile Forecasts Combination

Combining density forecasts has recently become a well-known practice for macroeconomic forecasting, see Aastveit et al. (2019) for a recent survey of the literature. The most common approach for combining predictive densities is to use weighted linear combinations of prediction models, evaluated using a type of scoring rule (e.g. Hall and Mitchell, 2007; Amisano and Giacomini, 2007; Jore et al., 2010; Hoogerheide et al., 2010; Kascha and Ravazzolo, 2010; Geweke and Amisano, 2011, 2012; Gneiting and Ranjan, 2013; Aastveit et al., 2014; Kapetanios et al., 2015; Ganics, 2017; Ganics et al., 2023). However, recent advances also include more complex combination approaches that allow for time-varying weights with possibly both learning and model set incompleteness (e.g. Billio et al., 2013; Casarin et al., 2015; Pettenuzzo and Ravazzolo, 2016; Del Negro et al., 2016; Aastveit et al., 2018; McAlinn and West, 2019; McAlinn et al., 2020; Aastveit et al., 2023).

Common to all the aforementioned approaches is that a single weight is attached to the entire predictive distribution for each model, assuming the predictive ability is constant across the various regions of the distribution. In the framework of quantile regression, applying the aforementioned approaches requires to (i) estimate the quantile regression, (ii) fit a distribution (skew t-distribution in Adrian et al., 2019) to obtain a smooth density forecast, and (iii) combine the K densities. In this paper we propose to directly combine the quantile forecasts obtained in step (i) and then fit a distribution to the resulting combined quantiles.

Suppose that a set of $k = 1, \ldots, K$ quantile forecasts distributions $\hat{y}_{t+h,q,k}$ for the same variable of interest y_t at horizon h are available. Standard density combination methods apply a unique combination weight to the entire predictive distribution, i.e.:

$$\underset{1\times Q}{y^c_{t+h}} = \underset{1\times K}{w_k}\ \underset{K\times Q}{\hat{y}_{t+h,q,k}} \tag{3}$$

where y^c_{t+h} denotes the resulting combined forecast for y, and $q = 1, \ldots, Q$ indicates the quantiles (or bins) of the predictive distribution. However, this procedure implicitly

overlooks superior forecast accuracy of some forecasts k over a specific region q of the distribution. Suppose indeed that a subset of this set is more accurate in predicting the mean (tails) of the distribution, while they perform poorly in the tails (mean). It would then be desirable to consider this heterogeneity when constructing the combined density i.e.:

$$\underset{1\times Q}{y^{c}_{t+h}} = diag\underset{Q\times K}{\left(w_{q,k}}\underset{K\times Q}{\hat{y}_{t+h,q,k}}\right) \tag{4}$$

Since here weights are quantile-specific, forecasts are now multiplied by a vector of combination weights instead of a scalar as in equation (3).

2.2.1 Evaluation of quantiles' forecast accuracy: the quantile scores

From our quantile ARDL regression (equation (1)) we obtain K (equal to 9 in our empirical application) forecasts for y_{t+h}, distributed over Q (set to 17 in our empirical application) quantiles. The purpose of this paper is to combine them taking into consideration the forecast accuracy at the quantile level. In order to do so, we need an evaluation method that helps us to discriminate not only the accuracy of the k^{th} forecast but also its accuracy at the q^{th} quantile. A common scoring rule for evaluating density forecasts is the Continuous Ranked Probability Score (CRPS). According to this score, the density forecast is evaluated by computing the distance at each point of the distribution to the realization. It is defined by:

$$CRPS = -\int_{-\infty}^{\infty}(F(y)_{t+h,q,k} - \mathbb{I}(F(y)_{t+h,q,k} \geq y_{t+h}))^2 dy \tag{5}$$

where $F(y)_{t+h,q,k}$ represents the CDF of forecast $f(y)_{t+h,q,k}$ and y_{t+h} the corresponding realization. The CRPS corresponds to the integral of the Brier scores for the probability forecasts at all real-valued thresholds (Matheson and Winkler, 1976; Hersbach, 2000; Gneiting and Raftery, 2007). While this score is the average "error" across the domain of the distribution, we follow Gneiting and Ranjan (2011) who propose the following quantile decomposition of the CRPS:

$$CRPS = \int_{0}^{1} QS_{t+h,k}(q)dq, \tag{6}$$

where $QS_{t+h,k}(q)$ is labeled the quantile score and defined as:

$$\begin{aligned} QS_{t+h,k}(q) &= \frac{1}{n-h+1}\sum_{t=m}^{m+n-h} QS_q(F^{-1}(y)_{t+h,q,k}, y_{t+h}), \text{ and} \\ QS_q(F^{-1}(y)_{t+h,q,k}, y_{t+h}) &= 2\Big(\mathbb{I}\{y_{t+h} \leq F^{-1}(y)_{t+h,q,k}\} - q\Big)(F^{-1}(y)_{t+h,q,k} - y_{t+h}) \end{aligned} \tag{7}$$

where m and n are defined by the in-sample and out-of-sample periods, $F^{-1}(y)_{t+h,q,k}$ is the value the inverse of the CDF of $f(y)_{t+h,q,k}$ taken at quantile q.

We would like to highlight a couple of properties of function QS_q in equation (7). First, notice that the closer q is to zero, the lower are the probabilities that $\hat{y}_{t+h,q,k}$ will have a value lower than y_{t+h}; at the same time, the closer q is to one, the lower are the probabilities that $\hat{y}_{t+h,q,k}$ will have a value greater than y_{t+h} (Laio and Tamea, 2007). The quantile score based on equation (7) therefore has a concave shape. Second, since QS_q is a measure of loss accuracy, the forecast that obtains the lowest QS_q curve is preferred to the other alternatives. Finally, as proven by Friederichs and Hense (2008) the CRPS-quantile decomposition $QS(q)$ is a proper scoring rule and Gneiting and Raftery (2007) also show that it encompasses the asymmetric loss score proposed by Giacomini and Komunjer (2005).

2.2.2 Quantile-specific combination weights

We propose to use the quantile scores to construct quantile-specific combination weights:

$$\underset{K\times Q}{w_{t+h,q,k}} = \frac{\sum_{t=m}^{m+n-h} 1/QS_{t,k}(q)}{\sum_{k=1}^{K}\sum_{t=m}^{m+n-h} 1/QS_{t,k}(q)} \tag{8}$$

where $t = m, \ldots, m+n-h$ denotes the forecast origins. $w_{t+h,q,k}$ is the matrix $K \times Q$ of combination weights for forecast k. The recursive weights are then a function of past performance of each model k known at the time the forecast is made (t). We need to impose that $w_{t+h,q,k} \geq 0$ and that:

$$\sum_{k=1}^{K} w_{t+h,q,k} = 1$$

The combined quantile forecast y^c_{t+h} is obtained by multiplying the matrix of combination weights computed according to equation (8) with the matrix of quantile forecasts:

$$y^c_{t+h} = \text{diag}(w_{t+h,q,k} \times \hat{y}_{t+h,q,k}) \tag{9}$$

The diagonal of this matrix corresponds to the match between the vector of weights for k model and the corresponding forecast from model k.

2.3 Predictive distributions

From the previous sections, we obtained the combined quantiles of the predictive distribution. In order to recover the full probability density function we fit the skew t-distribution

developed by Azzalini and Capitanio (2003), following Adrian et al. (2019):

$$f(y|\mu,\sigma,\alpha,\nu) = \frac{2}{\sigma} t\left(\frac{y-\mu}{\sigma}|\nu\right) T\left(\alpha\frac{y-\mu}{\sigma}\sqrt{\frac{\nu+1}{\nu+\frac{y-\mu}{\sigma}}}|\nu+1\right) \tag{10}$$

where f and $F(\cdot)$ respectively denote the PDF and CDF of the Student t-distribution. The four parameters of the distribution are the location μ, scale σ, fatness ν, and shape α. For each quarter, we choose the four parameters $\{\mu_t, \sigma_t, \alpha_t, \nu_t\}$ of the skewed t-distribution f to minimize the square distance between our estimated quantile function and the quantile function of the skew t-distribution $F^{-1}(q|\mu,\sigma,\alpha,\nu)$ from (10) to match the 10, 25, 75 and 90 percent quantiles:

$$\{\hat{\mu}_{t+h}, \hat{\sigma}_{t+h}, \hat{\alpha}_{t+h}, \hat{\nu}_{t+h}\} = \underset{\mu,\sigma,\alpha,\nu}{\operatorname{argmin}} \sum_q \left(\boldsymbol{x}_t'\boldsymbol{\beta}_q - F^{-1}(q,\mu,\sigma,\alpha,\nu)\right)^2 \tag{11}$$

where $\hat{\mu}_{t+h} \in \mathbb{R}$, $\hat{\sigma}_{t+h} \in \mathbb{R}^+$, $\hat{\alpha}_{t+h} \in \mathbb{R}$, and $\hat{\nu}_{t+h} \in \mathbb{Z}^+$.

We follow Mitchell et al. (2024), who suggest to fit a distribution at the last stage of forecasting, to reduce the impact of such approximation on the forecast accuracy. In contrast to alternative combination approaches, we therefore evaluate and combine quantile forecasts first, and only once they are combined we approximate the predictive distribution from the quantiles.[6]

2.4 Comparison with alternative combination approaches

In the empirical application, we will compare the predictive distribution from our quantile combination approach with the predictive distributions obtained from three alternative combination approaches.

2.4.1 Equal Weights

The first alternative combination approach is to allocate an equal combination weight, $w_k = 1/K$, to each of the predictive distributions from the K individual models. The combined predictive distribution is then the following:

$$f(y_{t+h}) = \sum_{k=1}^{K} w_k f(y)_{t+h,k} \tag{12}$$

[6]We have also explored constructing predictive densities from quantile forecasts using the non-parametric approach of Mitchell et al. (2024). The advantage of this approach is that it can flexibly accommodate various distributional shapes. A comparison between our combination approach and alternative linear combinations fitted on non-parametric distribution is briefly discussed in section 3.4 and detailed results are reported in Appendix A.3.

Combination weights $w_k = 1/K$ assure that the combined distribution is still a distribution since $0 \leq w_k \leq 1$ and $\sum_{k=1}^{K} w_k = 1$.

2.4.2 Optimal Weights

The second combination approach is the so called "Optimal Weights" proposed by Hall and Mitchell (2007) and Geweke and Amisano (2011). Here, combination weights are obtained by maximizing a logarithmic score function:

$$w_k = \frac{1}{T-h} \sum_{t=1}^{T-h} ln(f(y)_{t+1,k}) \quad s.t. \quad w_k > 0, \quad \sum_{k=1}^{K} w_k = 1 \tag{13}$$

which is known as the log predictive score. Given the size of K, the inference algorithm for w_k in Conflitti et al. (2015) is used. See also Hall and Mitchell (2007), Jore et al. (2010), and Geweke and Amisano (2010) for a discussion on the use of the log score as a ranking device for the forecast ability of different models.

2.4.3 Bayesian Model Averaging

The third combination approach is Bayesian Model Averaging (BMA, henceforth). Here, the individual predictive densities are combined into a composite-weighted predictive distribution $f(y_{t+h}|I_K)$, given by

$$f(y_{t+h}|I_K) = \sum_{k=1}^{K} F\left(M_k\right) f(y_{t+h}|k) \tag{14}$$

where $F\left(M_k\right)$ is the posterior probability of model k, based on the predictive likelihood for model k. Mitchell and Hall (2005) discuss the analogy of the log score in a frequentist framework to the log predictive likelihood in a Bayesian framework, and how it relates to the Kullback-Leibler divergence. See Hoeting et al. (1999) for a review on BMA.

2.5 Forecast evaluation

We measure the relative forecast accuracy using the CRPS and evaluate their calibration using Probability Integral Transforms (PITs) tests. In particular, we compare the forecasting performance of the various individual models and alternative combination approaches with versions of the CRPS that penalize the loss in accuracy at certain regions of the target distribution proposed by Gneiting and Ranjan (2011).

$$\text{emphCRPS}_{t+h,k} \int_0^1 QS_q(F_{t+h,k}^{-1}(q), y_{t+h})\nu(q)dq \tag{15}$$

where ν is a non-negative weight function on the unit interval. For a constant weight function, equation (15) reduces to the unweighted score (see equation (6)), reported as "uniform" in Table 1 in section 3.2. However, we are also considering scores that put an extra emphasis on specific regions of the distributions, such as the "centre" $\nu(q) = q(1-q)$, the "left tail", $\nu(q) = (1-q)^2$, the "right tail", $\nu(q) = (2q-1)^2$, both "tails", $\nu(q) = (2q-1)^2$ and an additional emphasis on both tails, labeled "heavy tails", $\nu(q) = (2q-1)^4$.

We also assess the overall fit of a forecast density by testing goodness-of-fit relative to the "true," but unobserved density using the PITs, i.e., the CDF of the forecast evaluated at the subsequent realization of GDP growth, see Diebold et al. (1998). The PITs summarize the properties of the densities and may help us judge whether the densities are biased in a particular direction and whether the width of the densities have been roughly correct on average. For correctly calibrated forecast densities, the PITs, at a minimum, should be uniformly distributed. We test for correct calibration by applying theRossi and Sekhposyan (2019) test and the four-raw-moment test by Knüppel (2015).

We would like to highlight that, since our main goal is to combine forecasts at the quantile level to obtain more accurate overall density forecasts, we use CRPS and PITs as our main measure of forecast accuracy and calibration, respectively. In case one is only interested in a specific part of the distribution, e.g. picking the forecast with the most accurate left tail, it would be more desirable to use the quantile score for that particular quantile as a measure of forecast accuracy. The emphasized CRPS, however, could be viewed as a middle point between these two measures: it evaluates the overall density forecast but also accounts for the researcher's cost function by penalizing more heavily the loss in a specific part of the distribution.

3 Empirical Application

In this section, we analyze the performance of our quantile combination approach for forecasting US real GDP growth using real time data. The main goal of the exercise is to examine the forecasting performance of our quantile combination approach, compare it to commonly used alternative combination approaches and analyze what are the most informative predictors for the various parts of the predictive distribution of GDP growth.

3.1 Data

A vast amount of research has shown that a variety of economic and financial variables contain predictive information about future economic recessions and downturns. In our application we will consider in total $K = 9$ different predictors.[7] These are leading

[7]In principle, we could have selected a larger number of forecasts to combine (i.e., a large K), as our quantile combination approach does not have explicit limitations in terms of scalability. However, we choose to use a relatively small number of models in our empirical exercise to facilitate the interpretation

indicators that cover a broad range of the macroeconomy and that earlier studies have found to be particularly useful for predicting GDP growth and recessions.

We include real-time vintages of the following variables: the National Financial Condition Index (NFCI)[8], the University of Michigan Consumer Sentiment Index (ICS), unemployment rate (U), a credit spread (CrSpread) that measures the difference between BAA corporate bond yield and the 10 year treasury yield, residential investment (ResInv), new housing permits (Permit), total (PCE) and durable personal consumption expenditures (PCEDG) and industrial production (INDPRO). We provide detailed information about the various series, including data source and data transformation, in Table (A.1) in the appendix. Our baseline data sample covers the period 1973Q1-2019Q4 and we take into account data revisions using real-time data (see Table A.1 for details).

3.2 Out-of-sample density forecasts for US GDP growth

Our data sample covers the period 1973Q1-2019Q4. The initial forecasts are estimated on the data period 1973Q1-1995Q2 (in-sample period of 89 data observations). The full recursive out-of-sample forecast evaluation period runs from 1995Q3-2019Q4 (97 observations). We report forecasts for two horizons: one quarter ahead ($H = 1$) and one year ahead ($H = 4$). These forecasts are based on models that are estimated on an expanding window.[9] Our models are all quarterly ARDL models, and we employ quarterly real-time data vintages that reflect the information available shortly after the release of national accounts data, typically between three to four weeks into the first month of a quarter. Our primary aim in this empirical application is to demonstrate the effectiveness of our quantile combination approach. This involves constructing density forecasts that incorporate the heterogeneity in accuracy across regions of the forecast distribution from multiple models. We do not specifically consider the fact that some of our predictors are available at frequencies higher than quarterly intervals. However, our quantile combination framework can easily be adapted to allow for nowcasting scenarios, incorporating models that account for the flow of data releases and the mixed-frequency nature of predictors. Finally, we follow Romer and Romer (2000) and Clark (2011) among many others, and use the second available estimate of GDP as the actual measure.[10]

We compare the forecasting performance of our quantile combination approach with

of results, particularly for the model-specific weights.

[8]Note that most earlier studies only use the current vintage of NFCI in a pseudo out-of-sample framework to form predictions of GDP growth. One exception is Amburgey and McCracken (2023b), who construct real-time vintages of the NFCI. They find additional gains in the predictive content of NFCI for quantiles of GDP growth, particularly leading up to recessions. In our empirical application, we use the real-time vintages constructed by Amburgey and McCracken (2023b).

[9]Results based on estimates using a rolling window of 89 quarters are shown in Appendix A.2 and, as discussed in section 3.4, are very close to the ones reported in Table 1 below.

[10]Our results are robust to the use of alternative benchmark measures as discussed and shown in section 3.3 and shown in Appendix A.5.

commonly used alternative combination approaches. Table 1 compares density forecasting accuracy from our quantile combination approach (labeled Q-comb) with equal weighting combination of linear forecasts (EQ), optimal weighting combination of linear forecasts (OPT) and Bayesian model averaging (BMA). We report relative density forecasting accuracy, measured by standard CRPS (labeled as "Uniform" in Table 1) and various CRPS versions that penalize loss accuracy at specific regions of the target distribution as defined in equation (15). The table reports averages over the evaluation periods.

Table 1: Average CRPS values with emphasis on specific regions of the distribution

one-quarter-ahead forecasts

Emphasis	$\nu(q)$	EQ	OPT	BMA	Q-comb
Uniform	$\nu(q) = 1$	0.408***	0.623***	0.681***	0.278
Centre	$\nu(q) = q(1-q)$	0.450***	0.684***	0.750***	0.054
Tails	$\nu(q) = (2q-1)^2$	0.308***	0.481***	0.525***	0.062
Right Tail	$\nu(q) = (2q-1)^2$	0.273***	0.778**	0.622***	0.084
Left Tail	$\nu(q) = (2q-1)^2$	0.647***	0.480***	0.672***	0.086
Heavy Tails	$\nu(q) = (2q-1)^4$	0.264***	0.418***	0.452***	0.028

one-year-ahead forecasts

Emphasis	$\nu(q)$	EQ	OPT	BMA	Q-comb
Uniform	$\nu(q) = q(1-q)$	0.523***	0.853***	0.898**	0.359
Centre	$\nu(q) = q(1-q)$	0.545***	0.892***	0.930	0.066
Tails	$\nu(q) = (2q-1)^2$	0.460***	0.756***	0.795***	0.093
Right Tail	$\nu(q) = (2q-1)^2$	0.358***	0.874**	0.847**	0.111
Left Tail	$\nu(q) = (2q-1)^2$	0.852	0.793***	0.906***	0.115
Heavy Tails	$\nu(q) = (2q-1)^4$	0.434***	0.708***	0.754***	0.046

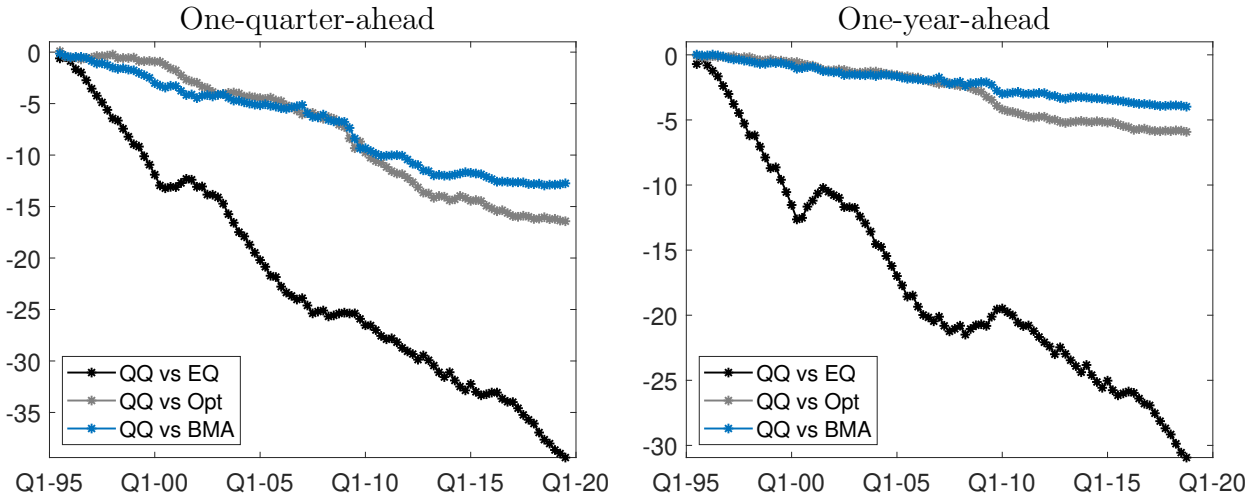
Note: The table reports average CRPS values with emphasis on specific regions of the distribution (see Eq. (15), for various forecast combination approaches. The alternative combination models EQ, OPT, BMA combines linear models, while Q-comb combines quantile models. For the alternative models, we report the relative performance compared to Q-comb. Thus, values > 1 denotes higher forecast accuracy than our quantile combination. Stars indicate significance levels for Diebold-Mariano test of Q-comb versus alternative approaches combinations.

The table reveals that our quantile combination approach outperforms the other alternative combination approaches for both forecasting horizons. This holds irrespective of using the CRPS or any quantile weighted version of the CRPS that emphasizes performance in either the centre, left or right tail of the distribution as forecast accuracy measure. This indicates that the relative gains in terms of forecasting performance from our combination approach are not specific to observations in a certain region of the dis-

tribution. Stars in the table represent the significance level of a Diebold-Mariano test for superior forecast ability of quantile combination versus the alternative linear combination approaches.

To address whether our improved out-of-sample forecast performance is limited to a specific time period or driven by some outliers, we report in Figure 1 the cumulative CRPS of the alternative combination approaches relative to our quantile combination. The measures are constructed so that a decrease in the relative value measures a relative improvement in the forecasting performance of our quantile combination compared to the alternative linear combination approach. While the various individual models show considerable instabilities in predictive performance over time, the performance from our quantile combination approach is far more robust, yielding a steady improvement over the various alternative combination approaches over the different time periods.

Figure 1: Cumulative CRPS of alternative combination approaches relative to the quantile combination for one-quarter and one-year ahead forecasts.



To better understand the superior predictive ability of our combination, we can inspect the combination weights' dynamics. Table 2 reports the combination weights at the end of the out-of-sample evaluation period, while Figures 2 and 3 show their evolution over time throughout the the out-of-sample evaluation period.

The results show that the combination weights vary, depending on the specific quantiles, time-period and forecasting horizon. NFCI and housing permits have a relatively high predictive power (reflected by a high weight) for the left tail and the center at the one quarter ahead forecasting horizon. However, predictive power for the one year ahead horizon and the right tail is low. Quite surprisingly, the credit spread, often used to forecast (financial) crises, is actually more informative for predicting the right tail. Residential investments has a relatively high predictive power for both the left and the (center) right tail at the one quarter ahead horizon, but is not so pronounced for the one year ahead horizon. Personal consumption expenditures (PCE and PCEDG), on the other had, do

not receive much weight for the short horizon forecasts, but are more important for forecasting one-year-ahead GDP. Moreover, Figures 2 and 3 show that there are periodically large fluctuations in the weights over time for all quantiles at both forecasting horizons.

Table 2: Combination weights at the end of evaluation sample

one-quarter-ahead forecasts

Q	NFCI	ICS	U	CrSpread	ResInv	PCE	PERMIT	PCEDG	INDPRO
0.10	0.1897	0.0864	0.0724	0.0757	0.1833	0.1246	0.0790	0.1006	0.0883
0.25	0.0822	0.1255	0.1051	0.0441	0.0681	0.0566	0.3147	0.0851	0.1185
0.50	0.2290	0.0699	0.0492	0.0718	0.1049	0.0529	0.2776	0.0598	0.0849
0.75	0.0433	0.0216	0.0288	0.2659	0.5152	0.0433	0.0299	0.0342	0.0177
0.90	0.1536	0.1527	0.0880	0.1204	0.1439	0.1185	0.0675	0.0832	0.0722

one-year-ahead forecasts

Q	NFCI	ICS	U	CrSpread	ResInv	PCE	PERMIT	PCEDG	INDPRO
0.10	0.1058	0.1125	0.1070	0.0750	0.0969	0.1413	0.1226	0.1410	0.0980
0.25	0.0802	0.0740	0.1159	0.1005	0.1889	0.1221	0.1695	0.0704	0.0785
0.50	0.0706	0.0880	0.1267	0.1481	0.0547	0.0821	0.0689	0.2587	0.1023
0.75	0.1623	0.0769	0.1112	0.1688	0.0912	0.1541	0.0964	0.0921	0.0470
0.90	0.0707	0.0720	0.0517	0.1042	0.0710	0.3702	0.0595	0.1390	0.0617

Note: The table shows the combination weight for each individual model for a specific quantile at the end of the out-of-sample evaluation period.

Our results support earlier findings that variables such as the NFCI (Adrian et al., 2019) and residential investments (Aastveit et al., 2019) are good predictors for the lower left tail of the GDP growth distribution. Interestingly, however, the results also show that there is scope for improving these forecasts by also adding information from additional predictors. In fact, models that include either the NFCI or residential investments do not consistently achieve the highest weights in our quantile combination. Note also that there seem to be clear gains from adding information from models using various predictors. Although there are some variables that clearly outperform other variables for specific quantiles and horizons, the weights do not seem to converge to either zero or one for any of the variables.

The results reveal a substantial heterogeneity in the predictive performance over the various quantiles and forecast horizons. None of the individual models perform equally well in forecasting over all the quantiles and for both forecasting horizons. The quantile combination approach proposed in this paper is a flexible way to account for the heterogeneity in accuracy over the predictive distribution of GDP growth and thereby lead to density forecast that are more accurate than forecasts from traditional combination methods.

Figure 2: Quantile combination weights over time for one-quarter-ahead forecasts for all $K = 9$ forecasting models.

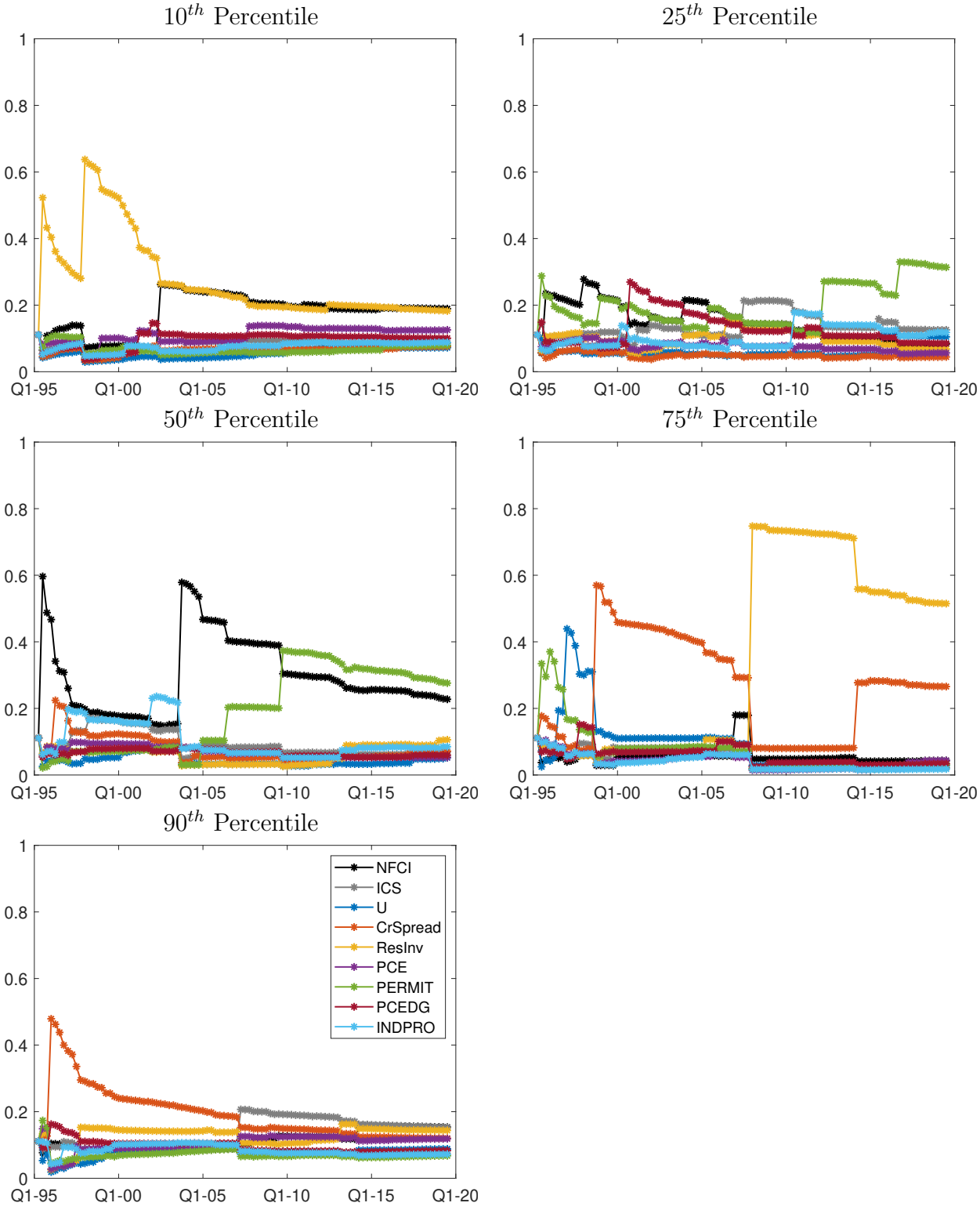
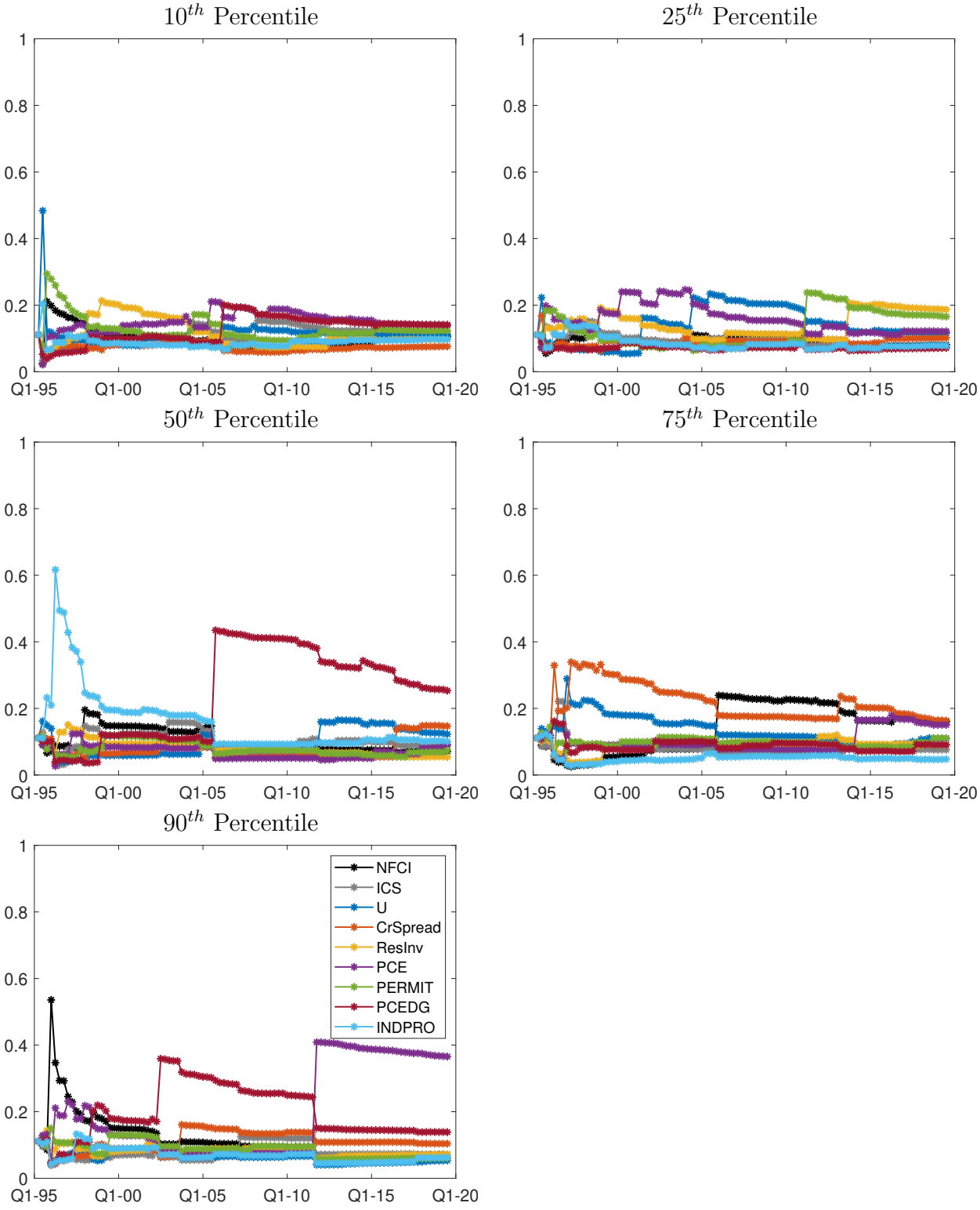


Figure 3: Quantile combination weights over time for one-year-ahead forecasts for all $K = 9$ forecasting models.



Finally, we assess whether the density forecasts from our quantile combination approach are well calibrated (or "probabilistically calibrated" as in the definition of Gneiting and Raftery, 2007). In Figure 4, we plot the empirical CDF of the PITs. Correctly calibrated forecasts would be displayed as an empirical CDF on the 45 degree line. The graph also shows 5% critical values calculated using bootstrapping techniques as in Rossi and Sekhposyan (2019). For completeness, the test statistics and corresponding critical values for the Kolmogorov-Smirnov and Cramer-von Mises (Rossi and Sekhposyan, 2019) and Knüppel (Knüppel, 2015) tests (using four moments) are reported in Table 3. The different tests all show that the density forecasts are well calibrated: according to the Kolmogorov-Smirnov and Cramer-von Mises tests, we fail to reject the null hypothesis of correct calibration at any usual significance level for both horizons. The Knüppel test confirms this.

Figure 4: Calibration of quantile-combined density forecast. Empirical CDF for PITs with the empirical 5% critical values calculated using Bootstrap following Rossi and Sekhposyan (2019).

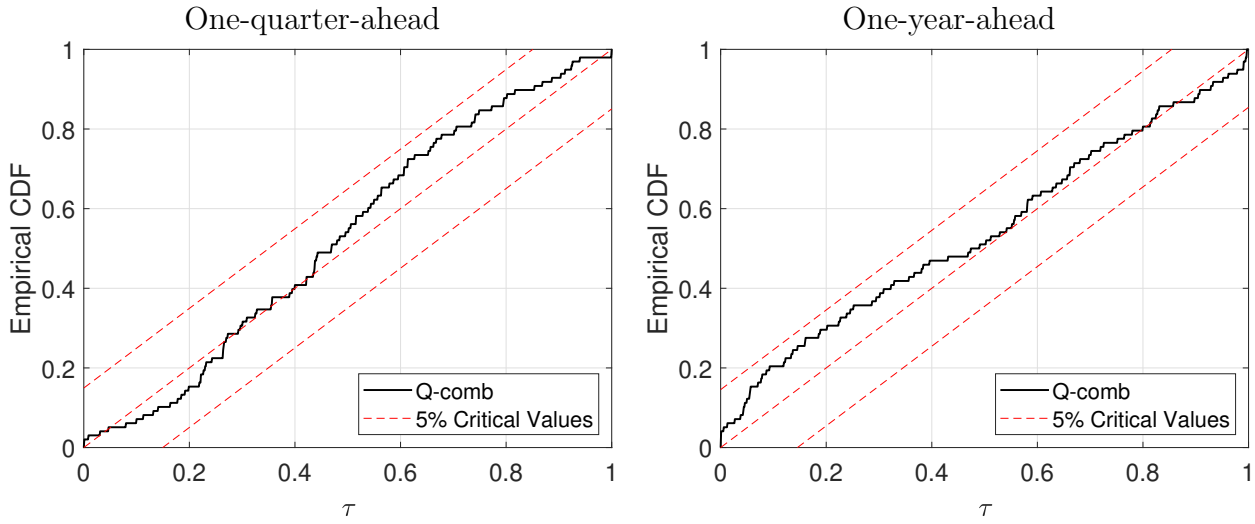


Table 3: Calibration tests statistics

one-quarter-ahead combined density forecasts

Calibration Tests	Statistic	Bootst. Critical Values at		
		10% s.l.	5% s.l.	1% s.l.
Kolmogorov–Smirnov	1.094	1.479	1.688	1.923
Cramér–von Mises	0.293	0.525	0.620	1.006
	p-values			
Knüppel NW	0.983			

one-year-ahead combined density forecasts

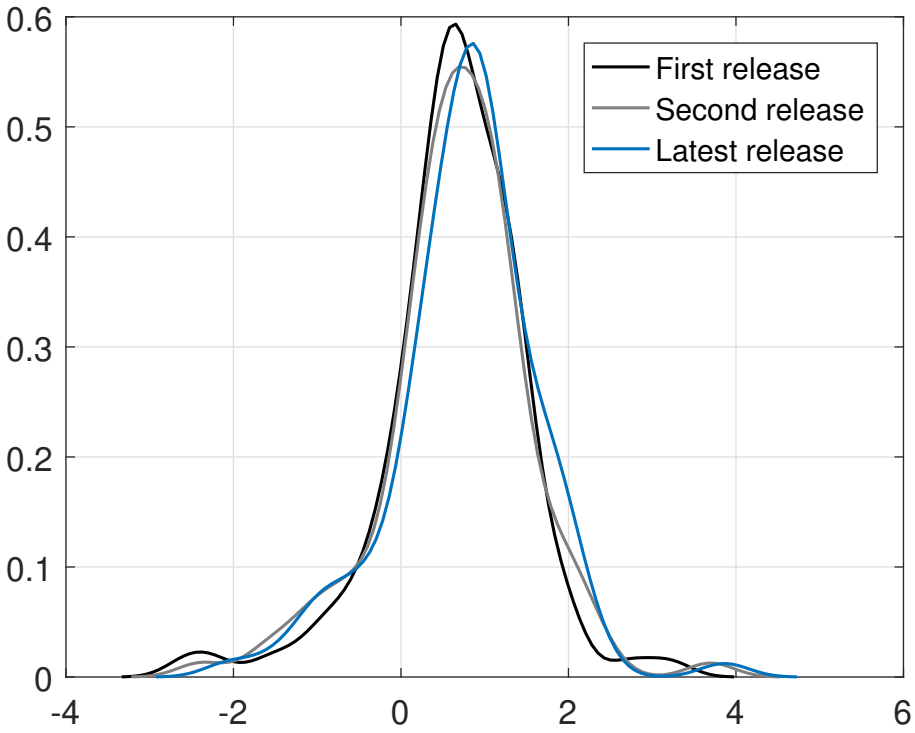
Calibration Tests	Statistic	Bootst. Critical Values at		
		10% s.l.	5% s.l.	1% s.l.
Kolmogorov–Smirnov	1.134	1.440	1.589	1.825
Cramér–von Mises	0.327	0.620	0.724	1.011
	p-values			
Knüppel NW	0.708			

Note: The null hypothesis of calibration is rejected for Kolmogorov–Smirnov and Cramér–von Mises if the test statistic is greater than the bootstrapped critical values following Rossi and Sekhposyan (2019); for Knüppel (2015) test rejects null hypothesis of calibration if the p-value displayed here is lower than significance level.

3.3 Alternative benchmark vintage

The choice of benchmark vintage for the "actual" measure of GDP is a key issue in any application using real-time vintage data (Croushore and Stark, 2001; Croushore, 2006). Stark and Croushore (2002) discuss various alternative benchmark data vintages and show that empirical results may depend on this choice. In our application, we follow Romer and Romer (2000) and Clark (2011), among many others, and use the second available estimate of GDP as the actual measure. However, in this section, we also consider two alternative benchmark vintages: the first release of GDP growth and the latest available vintage of GDP growth (the 2023Q3 release). In Figure 5, we plot the empirical distribution of the various GDP data vintages. The figure shows considerable differences between different GDP vintages, both in terms of the mean and the median. The differences are even more pronounced in the tails of the distributions. One may therefore expect that inference on quantile-specific forecasting performance may be even more sensitive to the choice of benchmark vintage than standard point or density forecasts.

Figure 5: Empirical distribution of GDP growth rate calculated from first, second and latest releases. Data until 2019Q4.



We study the implications for forecasting performance of our quantile combination approach when evaluating the forecasts against different vintages of GDP growth releases. Table 4 reports results for the comparison between our quantile combination approach and alternative combination approaches when evaluating the forecasts against the first release of GDP and the latest available release of GDP . As can be seen from the table, our quantile combination approach outperforms the other alternative combination approaches at both forecasting horizons, irrespective of using the CRPS or any quantile-weighted version of the CRPS that emphasizes performance in either the center, left, or right tail of the distribution as a forecast accuracy measure. In fact, the results are very similar to the baseline results in Table 1 and show that the superior performance of our quantile combination approach is robust to alternative benchmark vintages.[11]

In Tables 5 and 6, we report the combination weights at the end of the evaluation sample when using the first and the final vintage as the benchmark vintage, respectively. Interestingly, though, the model weights differ for the different vintages, compared to the weights in our baseline specification (Table 2). Based on the quantile-specific weights, the best-performing model for the various quantiles tends to differ in most cases. For instance, focusing on the one quarter ahead forecasts for the lower-left quantile, the consumer confidence survey model obtains the highest weight when using the latest vintage

[11]In appendix section A.5 we provide more detailed results, showing that the qunatile combination provide well-calibrated densities and yield a steady improvement over the various alternative combination approaches over time fore each of the alternative benchmark vintages.

Table 4: Average CRPS values with emphasis on specific regions of the distribution. First release of GDP and latest available release of GDP.

one-quarter-ahead forecasts

		first release				latest release			
Emphasis	$\nu(q)$	EQ	OPT	BMA	Q-comb	EQ	OPT	BMA	Q-comb
Uniform	$\nu(q) = 1$	0.420***	0.607***	0.676***	0.269	0.450***	0.678***	0.774***	0.322
Centre	$\nu(q) = q(1-q)$	0.469***	0.671***	0.746***	0.053	0.500***	0.750***	0.851***	0.063
Tails	$\nu(q) = (2q-1)^2$	0.314***	0.461***	0.513***	0.059	0.343***	0.522***	0.595***	0.072
Right Tail	$\nu(q) = (2q-1)^2$	0.288***	0.800*	0.651***	0.084	0.289***	0.786	0.724***	0.092
Left Tail	$\nu(q) = (2q-1)^2$	0.650***	0.444***	0.6300***	0.080	0.724***	0.556***	0.745***	0.105
Heavy Tails	$\nu(q) = (2q-1)^4$	0.273***	0.403***	0.450***	0.027	0.288***	0.444***	0.508***	0.032

one-year-ahead forecasts

		first release				latest release			
Emphasis	$\nu(q)$	EQ	OPT	BMA	Q-comb	EQ	OPT	BMA	Q-comb
Uniform	$\nu(q) = 1$	0.525***	0.841***	0.883**	0.338	0.569***	0.866***	0.892***	0.406
Centre	$\nu(q) = q(1-q)$	0.553***	0.887***	0.926*	0.063	0.595***	0.904***	0.938*	0.075
Tails	$\nu(q) = (2q-1)^2$	0.458***	0.737***	0.777***	0.087	0.51***	0.781***	0.805***	0.107
Right Tail	$\nu(q) = (2q-1)^2$	0.354***	0.881*	0.839**	0.104	0.391***	0.879**	0.844***	0.124
Left Tail	$\nu(q) = (2q-1)^2$	0.886	0.768***	0.879***	0.109	0.910	0.810***	0.898***	0.132
Heavy Tails	$\nu(q) = (2q-1)^4$	0.430***	0.694***	0.729***	0.043	0.477***	0.736***	0.757***	0.053

Note: The table reports average CRPS values with emphasis on specific regions of the distribution (see Eq. (15), for various forecast combination approaches. The alternative combination models EQ, OPT, BMA combines linear models, while Q-comb combines quantile models. For the alternative models, we report the relative performance compared to Q-comb. Thus, values > 1 denotes higher forecast accuracy than our quantile combination. Stars indicate significance levels for Diebold-Mariano test of Q-Comb versus alternative approaches combinations.

as the benchmark. This contrasts with our baseline results where NFCI was the most important predictor for the lower tail when using the second release of GDP as a benchmark. This example shows that inference based on forecasting performance from one specific part of the predictive distribution can be very sensitive to the specific choice of benchmark vintage. However, when measuring overall forecasting performance, our quantile combination approach seems robust to this, as it is flexible enough to adjust weights according to forecasting performance for different parts of the distribution depending on the specific choice of benchmark vintage.

Table 5: Combination weights at the end of evaluation sample - first release of GDP.

one-quarter-ahead forecasts

Q	NFCI	ICS	U	CrSpread	ResInv	PCE	PERMIT	PCEDG	INDPRO
0.10	0.1643	0.1225	0.0624	0.0808	0.1470	0.1093	0.1295	0.0897	0.0944
0.25	0.1118	0.1310	0.0860	0.0909	0.1318	0.1246	0.0891	0.0705	0.1643
0.50	0.3561	0.1316	0.2198	0.0440	0.0784	0.0575	0.0298	0.0339	0.0490
0.75	0.0813	0.0696	0.0439	0.0745	0.2232	0.0905	0.2386	0.0897	0.0888
0.90	0.0972	0.0891	0.0993	0.0937	0.1653	0.1823	0.0872	0.0914	0.0945

one-year-ahead forecasts

Q	NFCI	ICS	U	CrSpread	ResInv	PCE	PERMIT	PCEDG	INDPRO
0.10	0.0841	0.0919	0.1415	0.0999	0.0804	0.1044	0.1660	0.1206	0.1114
0.25	0.0900	0.0460	0.0442	0.0612	0.0654	0.2221	0.0727	0.3168	0.0817
0.50	0.0512	0.0237	0.0358	0.0380	0.0909	0.3670	0.0460	0.1261	0.2212
0.75	0.1479	0.1071	0.1145	0.1174	0.1056	0.0879	0.0982	0.1533	0.0682
0.90	0.0915	0.0627	0.1247	0.0771	0.1205	0.1376	0.0789	0.1190	0.1880

Note: The table shows the combination weight for each individual model for a specifc quantile at the end of the out-of-sample evaluation period.

Table 6: Combination weights at the end of evaluation sample - latest available release of GDP.

one-quarter-ahead forecasts

Q	NFCI	ICS	U	CrSpread	ResInv	PCE	PERMIT	PCEDG	INDPRO
0.1	0.0243	0.7941	0.0151	0.0371	0.0181	0.0305	0.0340	0.0102	0.0366
0.25	0.1310	0.0890	0.0810	0.1761	0.1051	0.1877	0.0580	0.1059	0.0662
0.50	0.0636	0.2290	0.2051	0.0714	0.0940	0.0931	0.0875	0.0659	0.0904
0.75	0.1482	0.0788	0.1024	0.1237	0.1347	0.1428	0.0859	0.0716	0.1118
0.90	0.0728	0.1150	0.0806	0.0733	0.1525	0.1541	0.0898	0.0706	0.1913

one-year-ahead forecasts

Q	NFCI	ICS	U	CrSpread	ResInv	PCE	PERMIT	PCEDG	INDPRO
0.1	0.0373	0.7002	0.0259	0.0279	0.0318	0.0351	0.0300	0.0768	0.0349
0.25	0.0867	0.0945	0.1306	0.1213	0.0969	0.0488	0.1843	0.1541	0.0828
0.50	0.1047	0.1221	0.0627	0.1140	0.2924	0.0485	0.0883	0.0755	0.0917
0.75	0.0622	0.0554	0.1963	0.0342	0.0521	0.0915	0.1110	0.1179	0.2795
0.90	0.0424	0.0904	0.0362	0.2085	0.0595	0.0927	0.0595	0.1455	0.2653

Note: The table shows the combination weight for each individual model for a specific quantile at the end of the out-of-sample evaluation period.

3.4 Robustness

We perform several robustness checks to evaluate the sensitivity of our results to the specifications chosen. The results are described in this section, with tables and figures in the Appendix.

3.4.1 Rolling window

We test whether our out-of-sample forecasting results are sensitive to using a recursive versus rolling window for model estimation. The baseline results are based on a recursive window, starting from 89 in-sample observations between 1973Q1-1995Q4 and expanding with each new forecast made. For the rolling window, we use a window size of 89, in line with the starting window for the baseline. The results can be found in Section A.2. From Table A.3 we can see that our quantile combination method still outperforms other combination methods for both standard and weighted evaluation criteria. Figure A.1 confirms that this is not time-dependent. Calibration tests show that the model is still well calibrated when using a rolling window for model estimation. As could be expected, the combination weights for the various quantiles and horizons are different from the baseline.

3.4.2 Non-parametric densities

We follow Mitchell et al. (2024) in constructing non-parametric densities using our quantile forecasts. As this only affects the combination evaluation, and not the quantile combination, the results on the weights are the same as the baseline results. The comparison of CRPS scores and the evaluation of calibration can be found in Section A.3. We can see from Table A.6 that our quantile combination method yields higher forecast accuracy than the alternative combination methods also for this specification. Table A.7 shows that the model is less well calibrated for the one-year-ahead horizon when applying non-parametric densities.

3.4.3 Including the COVID-19 pandemic

In order to evaluate the robustness of our conclusions when including the pandemic years in our sample, we increase the sample to 2023Q3. Results are shown in Section A.4 and are qualitatively similar.

4 Conclusion

In this paper, we propose a new forecast combination approach aimed at obtaining more accurate density forecasts for real GDP growth. The method assigns weights to the individual forecasts from the different indicators based on quantile scores as follows. First, individual forecasts are generated using quantile regression models. Subsequently, these individual forecasts are combined using a novel quantile combination approach, where each quantile of the combined density forecast is constructed as a weighted combination of the individual forecasts for the corresponding quantile. To address the heterogeneity in forecast accuracy across different models and parts of the distribution, we allocate

quantile-specific weights from each model using the quantile score introduced by Gneiting and Ranjan (2011).

In an empirical application, we demonstrate the usefulness of our novel quantile combination approach by forecasting real GDP growth for the Unites States for the period 1995Q1-2019Q4.

We show that density forecasts from our quantile combination approach outperform forecasts from commonly used combination approaches, including Bayesian Model Averaging, optimal combination of density forecasts as suggested by Hall and Mitchell (2007) and equal weights. The superior performance holds regardless of whether we use the standard CRPS or any quantile-weighted version of the CRPS that emphasizes accuracy in the centre, left or right tail of the distribution as the forecast accuracy measure. Importantly, the relative gains in forecasting performance from our combination approach are not specific to observations in a certain region of the distribution. This improved out-of-sample forecast performance is highly robust over time, showing a consistent and steady improvement compared to the various alternative combination approaches across different time periods. Therefore, the relative gain in terms of forecasting performance from our model is not limited to specific observations or sub-periods within our forecasting sample.

Additionally, while Adrian et al. (2019) argue that financial conditions are particularly informative about future downside macroeconomic risk, we demonstrate that quantile regressions incorporating additional variables such as residential investments, building permits and consumer confidence surveys yield more accurate forecasts for the lower left quantile of the GDP distribution compared to quantile regressions that solely include the NFCI. This finding suggests that there are other variables besides the NFCI that provide valuable information about future downside macroeconomic risk. Focusing exclusively on the NFCI may lead to overlooking crucial information from these other variables.

Furthermore, we document that data revisions have a greater impact on the tail of the distribution than on the middle, making inference on quantile-specific forecasting performance particularly sensitive to the choice of benchmark vintage compared to standard point or density forecasts. Despite this sensitivity, the forecasting performance of our quantile combination approach appears to be robust to the choice of benchmark vintage. The flexibility of our approach allows weights to adjust based on forecasting performance for different parts of the distribution, depending on the specific benchmark vintage selected.

Finally, our approach is flexible and easy to implement. While we demonstrate its usefulness by applying it to real GDP growth forecasting, its applicability extends far beyond this specific use case.

References

Aastveit, K. A., A. K. Anundsen, and E. I. Herstad (2019). Residential investment and recession predictability. *International Journal of Forecasting 35*(4), 1790–1799.

Aastveit, K. A., J. L. Cross, and H. K. van Dijk (2023). Quantifying time-varying forecast uncertainty and risk for the real price of oil. *Journal of Business & Economic Statistics 41*(2), 523–537.

Aastveit, K. A., K. R. Gerdrup, A. S. Jore, and L. A. Thorsrud (2014). Nowcasting GDP in real time: A density combination approach. *Journal of Business & Economic Statistics 32*, 48–68.

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

Aastveit, K. A., F. Ravazzolo, and H. K. van Dijk (2018). Combined density nowcasting in an uncertain economic environment. *Journal of Business & Economic Statistics 36*, 131–145.

Adrian, T., N. Boyarchenko, and D. Giannone (2019). Vulnerable growth. *American Economic Review 109*(4), 1263–89.

Amburgey, A. and M. W. McCracken (2023a). Growth-at-Risk is Investment-at-Risk. Working Papers 2023-020, Federal Reserve Bank of St. Louis.

Amburgey, A. and M. W. McCracken (2023b). On the real-time predictive content of financial condition indices for growth. *Journal of Applied Econometrics 38*(2), 137–163.

Amisano, G. and R. Giacomini (2007). Comparing density forecasts via weighted likelihood ratio tests. *Journal of Business & Economic Statistics 25*(2), 177–190.

Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(2), 367–389.

Billio, M., R. Casarin, F. Ravazzolo, and H. K. van Dijk (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics 177*, 213–232.

Brownlees, C. and A. B. Souza (2021). Backtesting global Growth-at-Risk. *Journal of Monetary Economics 118*(C), 312–330.

Carriero, A., T. E. Clark, and M. Marcellino (2022). Nowcasting tail risk to economic activity at a weekly frequency. *Journal of Applied Econometrics 37*(5), 843–866.

Casarin, R., S. Grassi, F. Ravazzolo, and H. K. van Dijk (2015). Parallel sequential Monte Carlo for efficient density combination: The DeCo MATLAB toolbox. *Journal of Statistical Software, Articles 68*, 1–30.

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010). Quantile and probability curves without crossing. *Econometrica 78*(3), 1093–1125.

Clark, T. E. (2011). Real-time density forecasts from bayesian vector autoregressions with stochastic volatility. *Journal of Business & Economic Statistics 29*(3), 327–341.

Clark, T. E., F. Huber, G. Koop, M. Marcellino, and M. Pfarrhofer (2023). Tail Forecasting With Multivariate Bayesian Additive Regression Trees. *International Economic Review 64*(3), 979–1022.

Coe, P. J. and S. P. Vahey (2020). Financial conditions and the risks to economic growth in the United States since 1875. CAMA Working Papers 2020-36, Centre for Applied Macroeconomic Analysis, Crawford School of Public Policy, The Australian National University.

Conflitti, C., C. De Mol, and D. Giannone (2015). Optimal combination of survey forecasts. *International Journal of Forecasting 31*(4), 1096–1103.

Croushore, D. (2006). *Forecasting with Real-Time Macroeconomic Data*, Volume 1 of *Handbook of Economic Forecasting*, Chapter 17, pp. 961–982. Elsevier.

Croushore, D. and T. Stark (2001). A real-time data set for macroeconomists. *Journal of Econometrics 105*(1), 111–130.

Del Negro, M., R. B. Hasegawa, and F. Schorfheide (2016). Dynamic prediction pools: An investigation of financial frictions and forecasting performance. *Journal of Econometrics 192*(2), 391–405.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review 39*(4), 863–83.

Diks, C., V. Panchenko, and D. van Dijk (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics 163*(2), 215–230.

Friederichs, P. and A. Hense (2008). A probabilistic forecast approach for daily precipitation totals. *Weather and Forecasting 23*(4), 659–673.

Ganics, G., B. Rossi, and T. Sekhposyan (2023). From fixed-event to fixed-horizon density forecasts: Obtaining measures of multihorizon uncertainty from survey density forecasts. *Journal of Money, Credit and Banking* (Forthcoming).

Ganics, G. A. (2017). Optimal density forecast combinations. Working papers, Banco de España.

Geweke, J. and G. Amisano (2010). Comparing and evaluating bayesian predictive distributions of asset returns. *International Journal of Forecasting 26*(2), 216–230.

Geweke, J. and G. Amisano (2011). Optimal prediction pools. *Journal of Econometrics 164*(1), 130–141.

Geweke, J. and G. G. Amisano (2012). Prediction with misspecified models. *The American Economic Review 102*, 482–486.

Giacomini, R. and I. Komunjer (2005). Evaluation and combination of conditional quantile forecasts. *Journal of Business & Economic Statistics 23*(4), 416–431.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association 102*(477), 359–378.

Gneiting, T. and R. Ranjan (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics 29*(3), 411–422.

Gneiting, T. and R. Ranjan (2013). Combining predictive distributions. *Electronic Journal of Statistics 7*, 1747–1782.

Hall, S. G. and J. Mitchell (2007). Combining density forecasts. *International Journal of Forecasting 23*(1), 1–13.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting 15*(5), 559–570.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 382–401.

Hoogerheide, L., R. Kleijn, F. Ravazzolo, H. K. Van Dijk, and M. Verbeek (2010). Forecast accuracy and economic gains from Bayesian model averaging using time-varying weights. *Journal of Forecasting 29*, 251–269.

Jore, A. S., J. Mitchell, and S. P. Vahey (2010). Combining forecast densities from vars with uncertain instabilities. *Journal of Applied Econometrics 25*(4), 621–634.

Kapetanios, G., J. Mitchell, S. Price, and N. Fawcett (2015). Generalised density forecast combinations. *Journal of Econometrics 188*(1), 150–165.

Kascha, C. and F. Ravazzolo (2010). Combining inflation density forecasts. *Journal of forecasting 29*(1-2), 231–250.

Knüppel, M. (2015). Evaluating the calibration of multi-step-ahead density forecasts using raw moments. *Journal of Business & Economic Statistics 33*(2), 270–281.

Koenker, R. and G. Bassett Jr (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33–50.

Laio, F. and S. Tamea (2007). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences 11*(4), 1267–1277.

Liu, W. and E. Moench (2016). What predicts us recessions? *International Journal of Forecasting 32*(4), 1138–1150.

Manzan, S. (2015). Forecasting the distribution of economic variables in a data-rich environment. *Journal of Business & Economic Statistics 33*(1), 144–164.

Marcellino, M. (2006). Leading indicators. In G. Elliott, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 1, pp. 879–960. Amsterdam: Elsevier.

Matheson, J. E. and R. L. Winkler (1976). Scoring rules for continuous probability distributions. *Management Science 22*(10), 1087–1096.

McAlinn, K., K. A. Aastveit, J. Nakajima, and M. West (2020). Multivariate Bayesian Predictive Synthesis in Macroeconomic Forecasting. *Journal of the American Statistical Association*. Forthcoming.

McAlinn, K. and M. West (2019). Dynamic bayesian predictive synthesis in time series forecasting. *Journal of Econometrics 210*(1), 155–169.

Mitchell, J. and S. G. Hall (2005). Evaluating, comparing and combining density forecasts using the klic with an application to the bank of england and niesr 'fan' charts of inflation. *Oxford Bulletin of Economics and Statistics 67*(s1), 995–1033.

Mitchell, J., A. Poon, and D. Zhu (2024). Constructing density forecasts from quantile regressions: Multimodality in macrofinancial dynamics. *Journal of Applied Econometrics* (Forthcoming).

Opschoor, A., D. Van Dijk, and M. van der Wel (2017). Combining density forecasts using focused scoring rules. *Journal of Applied Econometrics 32*(7), 1298–1313.

Pettenuzzo, D. and F. Ravazzolo (2016). Optimal portfolio choice under decision-based model combinations. *Journal of Applied Econometrics 31*, 1312–1332.

Reichlin, L., G. Ricco, and T. Hasenzagl (2020). Financial variables as predictors of real growth vulnerability. Discussion Papers 05/2020, Deutsche Bundesbank.

Romer, C. D. and D. H. Romer (2000). Federal reserve information and the behavior of interest rates. *American Economic Review 90*(3), 429–457.

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

Stark, T. and D. Croushore (2002). Forecasting with a real-time data set for macroeconomists. *Journal of Macroeconomics 24*(4), 507–531.

A Online Appendix

A.1 Description of data series

Table A.1: Description of data series

Label	Transf	Period	Real-Time	Description	Source
rgdp	Δln	59:Q1-19:Q4	91Q4-19Q4	Real GDP growth, sa	AL
NFCI	level	73:Q1-19:Q4	88Q1-19Q4	National Financial Conditions Index	CF and AM
ICS	level-100	60:Q1-19:Q4	98Q3-19Q4	Consumer Sentiment Index	AL
U	Δ log	48:Q1-19:Q4	65Q4-19Q4	Unemployment rate	AL
CrSpread	Level	53:Q1-19:Q4	none	Credit Spread: BAA corporate bond yield - 10-year treasury	F
ResInv	Δ%	47:Q2-19:Q4	65Q4-19Q4	Real Gross Private Domestic Investment: Fixed Investment: Residential	AL
PCE	Δ%	59:Q1-19Q4	79Q4-19:Q4	Personal Consumption Expenditures, SA, Annual Rate	AL
Permit	Δ%	60:Q1-19:Q4	99:Q3-19:Q4	New Privately-Owned Housing Units Authorized (Total), SA Annual Rate	AL
PCEDG	Δ%	59:Q1-19:Q4	79Q4-19:Q4	Personal Consumption Expenditures: Durable Goods, SA Annual Rate	AL
INDPRO	Δ%	59Q1:19Q4	70Q1:19Q4	Industrial Production: Total Index, SA	AL

Notes: Sources abbreviated as "F" denotes Federal Reserve Economic Data (FRED), as "AL" denotes Federal Reserve Economic Real-Data (ALFRED) dataset, as "CF" denotes Federal Reserve of Chicago and "AM" denotes Amburgey and McCracken (2023b).

A.2 Rolling window results

Table A.3: Average CRPS values with emphasis on specific regions of the distribution - using a rolling window of 89 observations.

one-quarter-ahead forecasts

Emphasis	$\nu(q)$	EQ	OPT	BMA	Q-comb
Uniform	$\nu(q) = 1$	0.428***	0.670***	0.785***	0.292
Centre	$\nu(q) = q(1-q)$	0.471***	0.731***	0.864***	0.057
Tails	$\nu(q) = (2q-1)^2$	0.328***	0.524***	0.617***	0.066
Right Tail	$\nu(q) = (2q-1)^2$	0.286***	0.772**	0.752***	0.088
Left Tail	$\nu(q) = (1-q)^2$	0.679***	0.548***	0.746***	0.091
Heavy Tails	$\nu(q) = (2q-1)^4$	0.283***	0.455***	0.536***	0.030

one-year-ahead forecasts

Emphasis	$\nu(q)$	EQ	OPT	BMA	Q-comb
Uniform	$\nu(q) = 1$	0.542***	0.867***	0.932**	0.372
Centre	$\nu(q) = q(1-q)$	0.570***	0.908***	0.972	0.069
Tails	$\nu(q) = (2q-1)^2$	0.480***	0.776***	0.829***	0.097
Right Tail	$\nu(q) = (2q-1)^2$	0.374***	0.899*	0.892***	0.116
Left Tail	$\nu(q) = (1-q)^2$	0.881	0.804***	0.930***	0.119
Heavy Tails	$\nu(q) = (2q-1)^4$	0.453***	0.727***	0.787***	0.048

Note: The table reports average CRPS values with emphasis on specific regions of the distribution (see Eq. (15), for various forecast combination approaches. The alternative combination models EQ, OPT, BMA combines linear models, while Q-comb combines quantile models. For the alternative models, we report the relative performance compared to Q-comb. Thus, values > 1 denotes higher forecast accuracy than our quantile combination. Stars indicate significance levels for Diebold-Mariano test of Q-comb versus alternative approaches combinations.

Figure A.1: Cumulative CRPS of the alternative combination approaches relative to quantile combination for one-quarter and one-year ahead forecasts - using a rolling window of 89 observations.

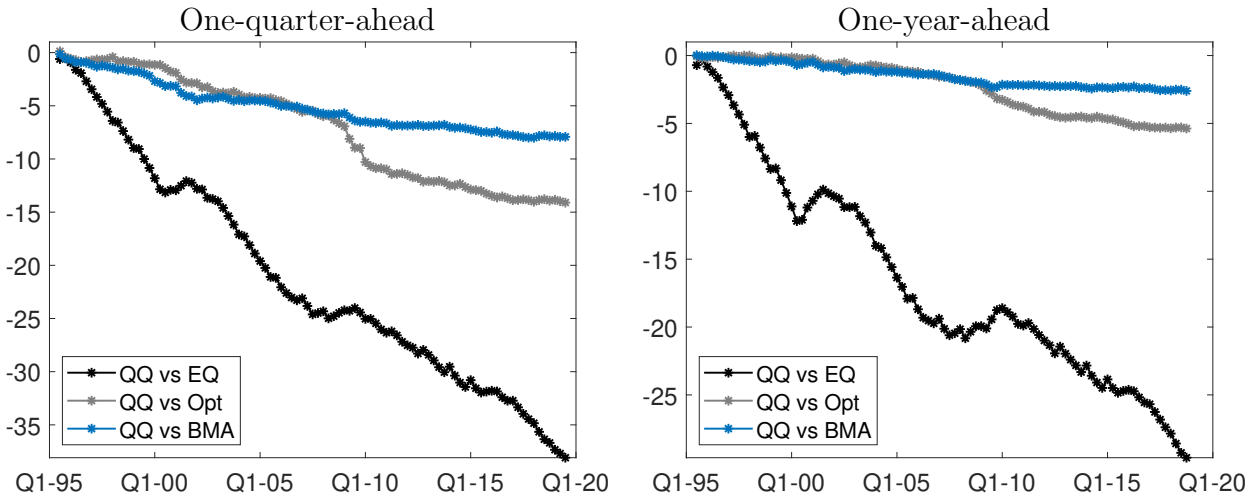


Table A.4: Combination weights at the end of evaluation sample - using a rolling window of 89 observations

one-quarter-ahead forecasts

Q	NFCI	ICS	U	CrSpread	ResInv	PCE	PERMIT	PCEDG	INDPRO
0.10	0.1047	0.0749	0.1016	0.0583	0.0892	0.1115	0.1170	0.0933	0.2496
0.25	0.1041	0.1015	0.2280	0.0942	0.0551	0.0582	0.2095	0.0769	0.0724
0.50	0.1077	0.1001	0.0990	0.0902	0.0922	0.1518	0.0873	0.0769	0.1947
0.75	0.1079	0.0952	0.2141	0.0618	0.0894	0.0646	0.1106	0.0733	0.1832
0.90	0.0641	0.0791	0.1047	0.1977	0.0695	0.1174	0.2034	0.0814	0.0828

one-year-ahead forecasts

Q	NFCI	ICS	U	CrSpread	ResInv	PCE	PERMIT	PCEDG	INDPRO
0.10	0.0568	0.0534	0.1278	0.0713	0.0924	0.0651	0.0588	0.2634	0.2111
0.25	0.1656	0.0501	0.0465	0.0799	0.0581	0.0981	0.0866	0.0724	0.3427
0.50	0.0291	0.0186	0.8484	0.0300	0.0099	0.0108	0.0185	0.0213	0.0135
0.75	0.1930	0.0472	0.3232	0.0523	0.0615	0.0549	0.0991	0.1115	0.0572
0.90	0.0832	0.0849	0.0853	0.0616	0.0612	0.0607	0.0549	0.4382	0.0699

Note: The table shows the combination weight for each individual model for a specific quantile at the end of the out-of-sample evaluation period.

Figure A.2: Quantile combination weights over time for one-quarter-ahead forecasts for all $K = 9$ forecasting models - using a rolling window of 89 observations.

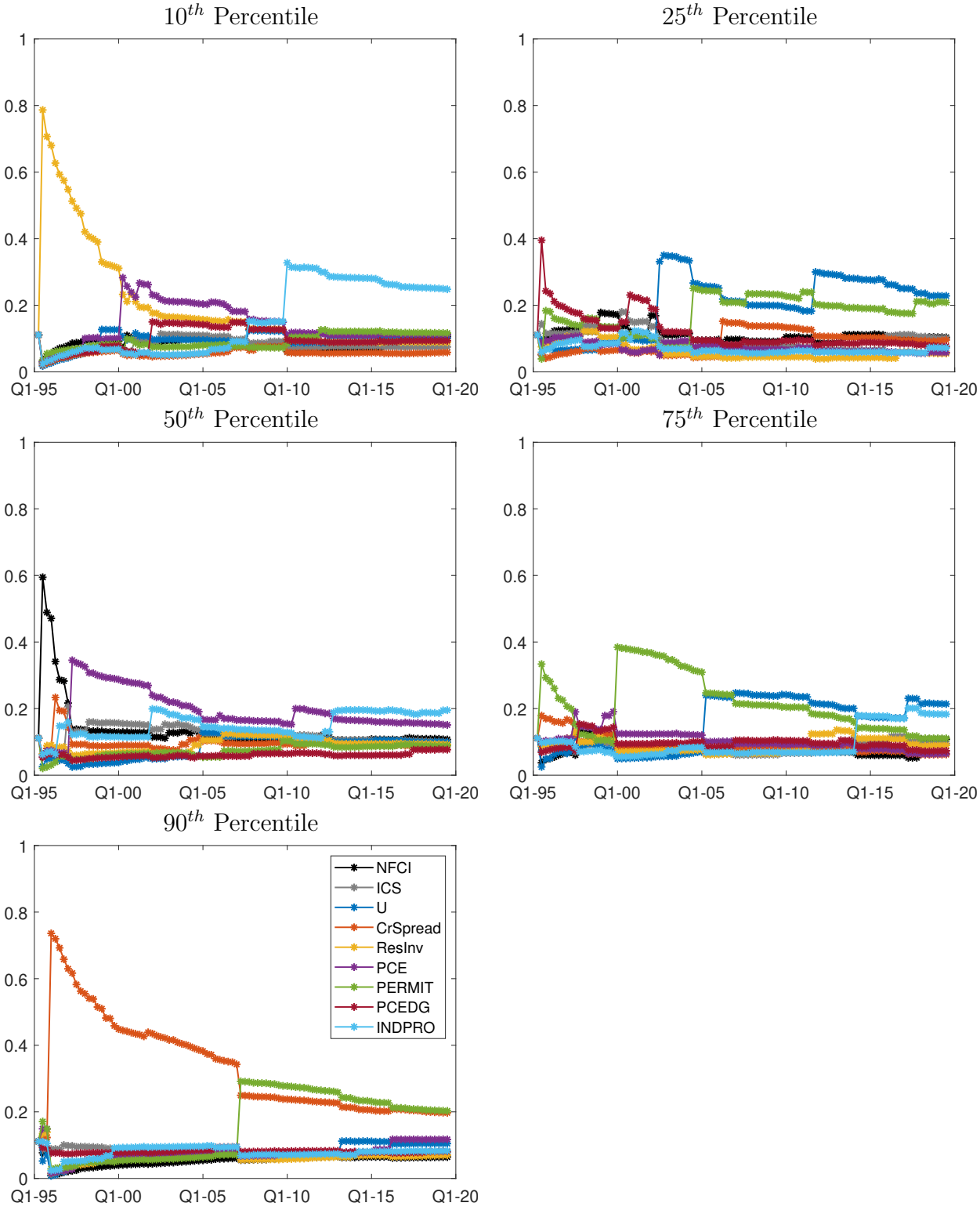


Figure A.3: Quantile combination weights over time for one-year-ahead forecasts for all $K = 9$ forecasting models- using a rolling window of 89 observations.

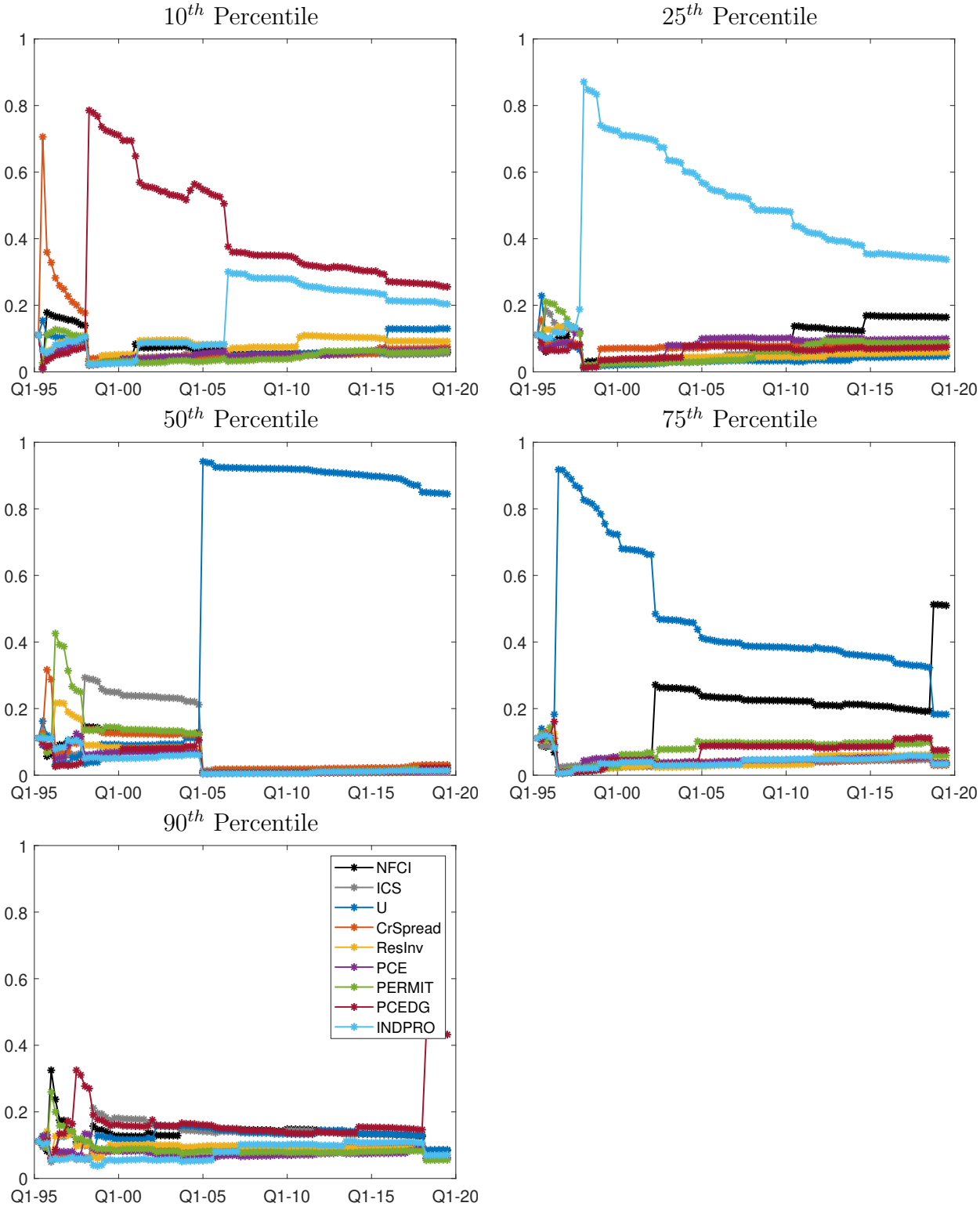


Figure A.4: Calibration of quantile-combined density forecast. Empirical CDF for PITs with the empirical 5% critical values calculated using Bootstrap following Rossi and Sekhposyan (2019) - using a rolling window of 89 observations.

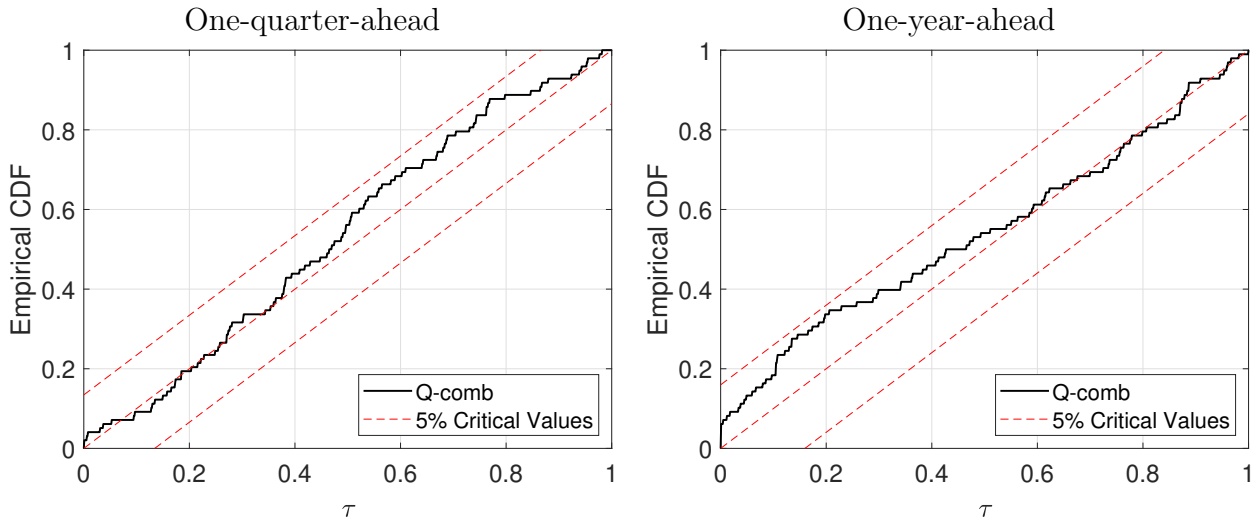


Table A.5: Calibration tests statistics for quantile-combined forecasts - using a rolling window of 89 observations.

one-quarter-ahead combined density forecasts

Calibration Tests	Statistic	Bootst. Critical Values at		
		10% s.l.	5% s.l.	1% s.l.
Kolmogorov–Smirnov	1.075	1.333	1.472	1.810
Cramér–von Mises	0.258	0.445	0.584	0.816
	p-values			
Knüppel NW	0.972			

one-year-ahead combined density forecasts

Calibration Tests	Statistic	Bootst. Critical Values at		
		10% s.l.	5% s.l.	1% s.l.
Kolmogorov–Smirnov	1.395	1.579	1.743	2.054
Cramér–von Mises 0.419	0.674	0.879	1.348	
	p-values			
Knüppel NW	0.846			

Note: The null hypothesis of calibration is rejected for Kolmogorov–Smirnov and Cramér–von Mises if the test statistic is greater than the bootstrapped critical values following Rossi and Sekhposyan (2019); for Knüppel (2015) test rejects null hypothesis of calibration if the p-value displayed here is lower than significance level.

A.3 Constructing the density forecast non-parametrically following Mitchell et al. (2024)

Table A.6: Average CRPS values with emphasis on specific regions of the distribution - quantiles fitted on a non-parametric distribution.

one-quarter-ahead forecasts

Emphasis	$\nu(q)$	EQ	OPT	BMA	Q-comb
Uniform	$\nu(q) = 1$	0.446***	0.677***	0.744***	0.319
Centre	$\nu(q) = q(1-q)$	0.492***	0.738***	0.816***	0.062
Tails	$\nu(q) = (2q-1)^2$	0.338***	0.526***	0.577***	0.071
Right Tail	$\nu(q) = (2q-1)^2$	0.289***	0.793**	0.662***	0.092
Left Tail	$\nu(q) = (2q-1)^2$	0.710***	0.551***	0.752***	0.103
Heavy Tails	$\nu(q) = (2q-1)^4$	0.288***	0.451***	0.500***	0.032

one-year-ahead forecasts

Uniform	$\nu(q) = 1$	0.564***	0.861***	0.892**	0.403
Centre	$\nu(q) = q(1-q)$	0.595***	0.904***	0.938	0.075
Tails	$\nu(q) = (2q-1)^2$	0.500***	0.766***	0.795***	0.105
Right Tail	$\nu(q) = (2q-1)^2$	0.385***	0.871**	0.836**	0.122
Left Tail	$\nu(q) = (1-q)^2$	0.910	0.815***	0.904***	0.132
Heavy Tails	$\nu(q) = (2q-1)^4$	0.468***	0.722***	0.754***	0.052

Note: The table reports average CRPS values with emphasis on specific regions of the distribution (see Eq. (15), for various forecast combination approaches. The alternative combination models EQ, OPT, BMA combines linear models, while Q-comb combines quantile models. For the alternative models, we report the relative performance compared to Q-comb. Thus, values > 1 denotes higher forecast accuracy than our quantile combination. Stars indicate significance levels for Diebold-Mariano test of Q-comb versus alternative approaches combinations.

Figure A.5: Cumulative CRPS of the alternative combination approaches relative to quantile combination for one-quarter and one-year ahead forecasts - quantiles fitted on a non-parametric distribution.

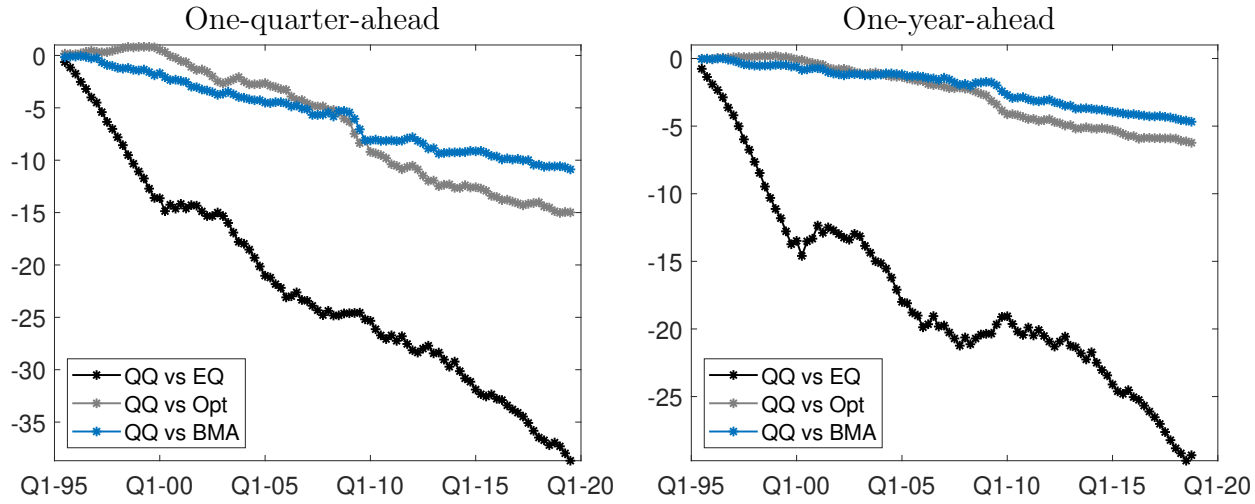


Figure A.6: Calibration of quantile-combined density forecast. Empirical CDF for PITs with the empirical 5% critical values calculated using Bootstrap following Rossi and Sekhposyan (2019) - quantiles fitted on a non-parametric distribution.

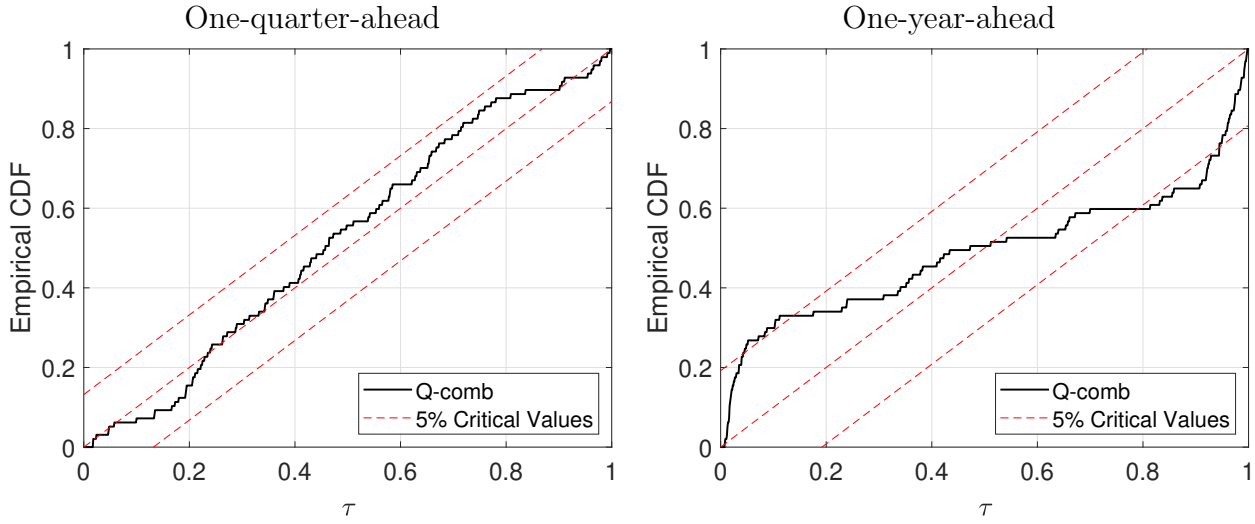


Table A.7: Calibration tests statistics for quantile-combined forecasts fitted on a non-parametric distribution.

one-quarter-ahead combined density forecasts

Calibration Tests	Statistic	Bootst. Critical Values at		
		10% s.l.	5% s.l.	1% s.l.
Kolmogorov–Smirnov	0.949	1.303	1.410	1.728
Cramér–von Mises	0.221	0.441	0.560	0.845
	p-values			
Knüppel NW	0.159			

one-year-ahead combined density forecasts

Calibration Tests	Statistic	Bootst. Critical Values at		
		10% s.l.	5% s.l.	1% s.l.
Kolmogorov–Smirnov	2.536	1.890	2.172	3.2884
Cramér–von Mises	1.895	0.999	1.301	2.684
	p-values			
Knüppel NW	0.021			

Note: The null hypothesis of calibration is rejected for Kolmogorov–Smirnov and Cramér–von Mises if the test statistic is greater than the bootstrapped critical values following Rossi and Sekhposyan (2019); for Knüppel (2015) test rejects null hypothesis of calibration if the p-value displayed here is lower than significance level.

A.4 Including the COVID-19 Pandemic

Table A.8: Average CRPS values with emphasis on specific regions of the distribution. Combined quantiles fitted to a skew t-distribution - sample up to 2023Q3, including the COVID-19 pandemic.

one-quarter-ahead forecasts

Emphasis	$\nu(q)$	EQ	OPT	BMA	Q-comb
Uniform	$\nu(q) = 1$	0.544***	0.633*	0.704**	0.455
Centre	$\nu(q) = q(1-q)$	0.574***	0.664	0.739*	0.085
Tails	$\nu(q) = (2q-1)^2$	0.459***	0.543*	0.608**	0.113
Right Tail	$\nu(q) = (2q-1)^2$	0.391***	0.676*	0.648*	0.138
Left Tail	$\nu(q) = (2q-1)^2$	0.777	0.564	0.716***	0.146
Heavy Tails	$\nu(q) = (2q-1)^4$	0.426***	0.505**	0.567**	0.055

one-year-ahead forecasts

Emphasis	$\nu(q)$	EQ	OPT	BMA	Q-comb
Uniform	$\nu(q) = 1$	0.689	0.825**	0.92**	0.584
Centre	$\nu(q) = q(1-q)$	0.707	0.848**	0.946*	0.106
Tails	$\nu(q) = (2q-1)^2$	0.639	0.768***	0.855***	0.159
Right Tail	$\nu(q) = (2q-1)^2$	0.510***	0.816**	0.879**	0.182
Left Tail	$\nu(q) = (2q-1)^2$	0.990*	0.808***	0.926**	0.189
Heavy Tails	$\nu(q) = (2q-1)^4$	0.611*	0.734***	0.816***	0.080

Note: The table reports average CRPS values with emphasis on specific regions of the distribution (see Eq. (15), for various forecast combination approaches. The alternative combination models EQ, OPT, BMA combines linear models, while Q-comb combines quantile models. For the alternative models, we report the relative performance compared to Q-comb. Thus, values > 1 denotes higher forecast accuracy than our quantile combination. Stars indicate significance levels for Diebold-Mariano test of Q-comb versus alternative approaches combinations.

Figure A.7: Cumulative CRPS of the alternative combination approaches relative to quantile combination for one-quarter and one-year ahead forecasts - sample up to 2023Q3, including the COVID-19 pandemic.

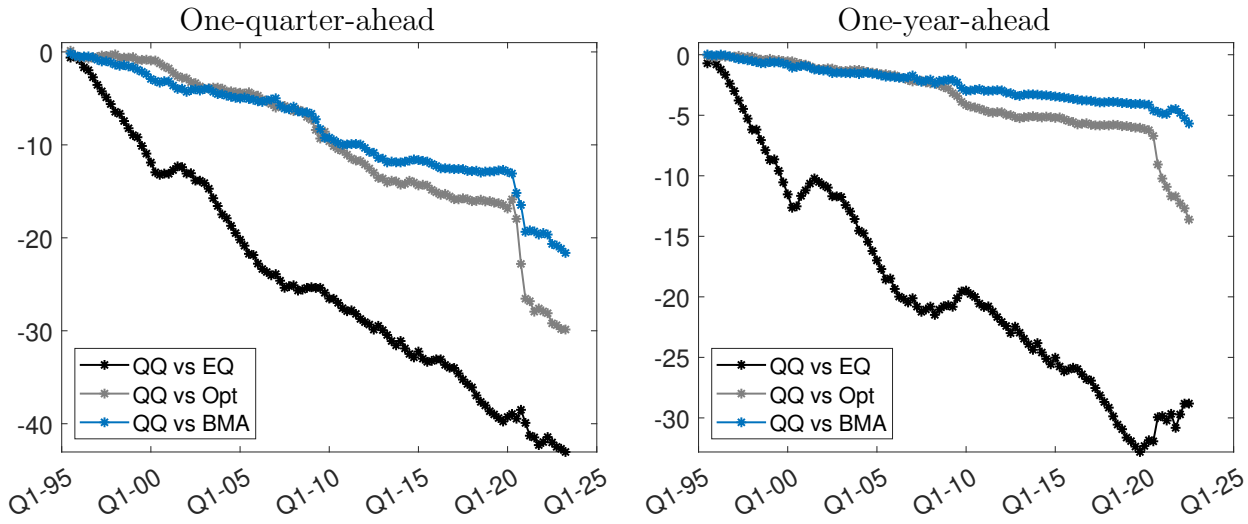


Table A.9: Combination weights at the end of evaluation sample - sample up to 2023Q3, including the COVID-19 pandemic.

one-quarter-ahead forecasts

Q	NFCI	ICS	U	CrSpread	ResInv	PCE	PERMIT	PCEDG	INDPRO
0.10	0.1765	0.1168	0.0683	0.0740	0.1704	0.1234	0.0765	0.1095	0.0845
0.25	0.0832	0.1259	0.1043	0.0469	0.0684	0.0557	0.3005	0.1013	0.1140
0.50	0.2224	0.0694	0.0525	0.0734	0.1057	0.0600	0.2704	0.0618	0.0843
0.75	0.0482	0.0225	0.0291	0.2538	0.5012	0.0423	0.0510	0.0334	0.0186
0.90	0.1494	0.1500	0.0884	0.1201	0.1457	0.1165	0.0718	0.0843	0.0739

one-year-ahead forecasts

Q	NFCI	ICS	U	CrSpread	ResInv	PCE	PERMIT	PCEDG	INDPRO
0.10	0.1051	0.1110	0.1217	0.0753	0.0981	0.1350	0.1217	0.1369	0.0950
0.25	0.0818	0.0785	0.1184	0.1021	0.1847	0.1208	0.1639	0.0711	0.0786
0.50	0.0705	0.0883	0.1237	0.1462	0.0540	0.0911	0.0731	0.2516	0.1015
0.75	0.1600	0.0764	0.1104	0.1627	0.0908	0.1502	0.1107	0.0907	0.0482
0.90	0.0716	0.0722	0.0586	0.1033	0.0717	0.3620	0.0604	0.1374	0.0628

Note: The table shows the combination weight for each individual model for a specific quantile at the end of the out-of-sample evaluation period.

Figure A.8: Quantile combination weights over time for one-quarter-ahead forecasts for all $K = 9$ forecasting models - sample up to 2023Q3, including the COVID-19 pandemic.

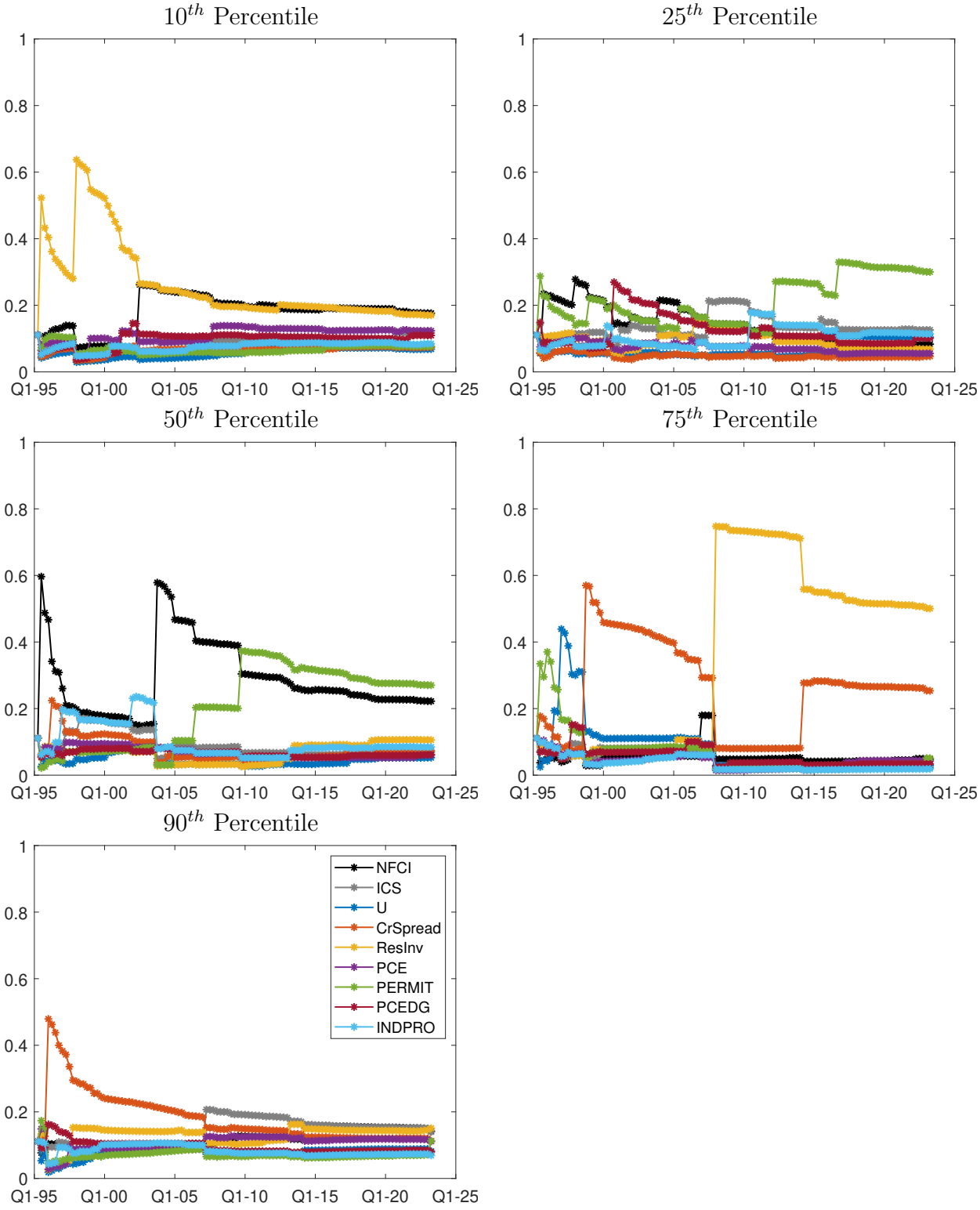


Figure A.9: Quantile combination weights over time for one-year-ahead forecasts for all $K = 9$ forecasting models - sample up to 2023Q3, including the COVID-19 pandemic.

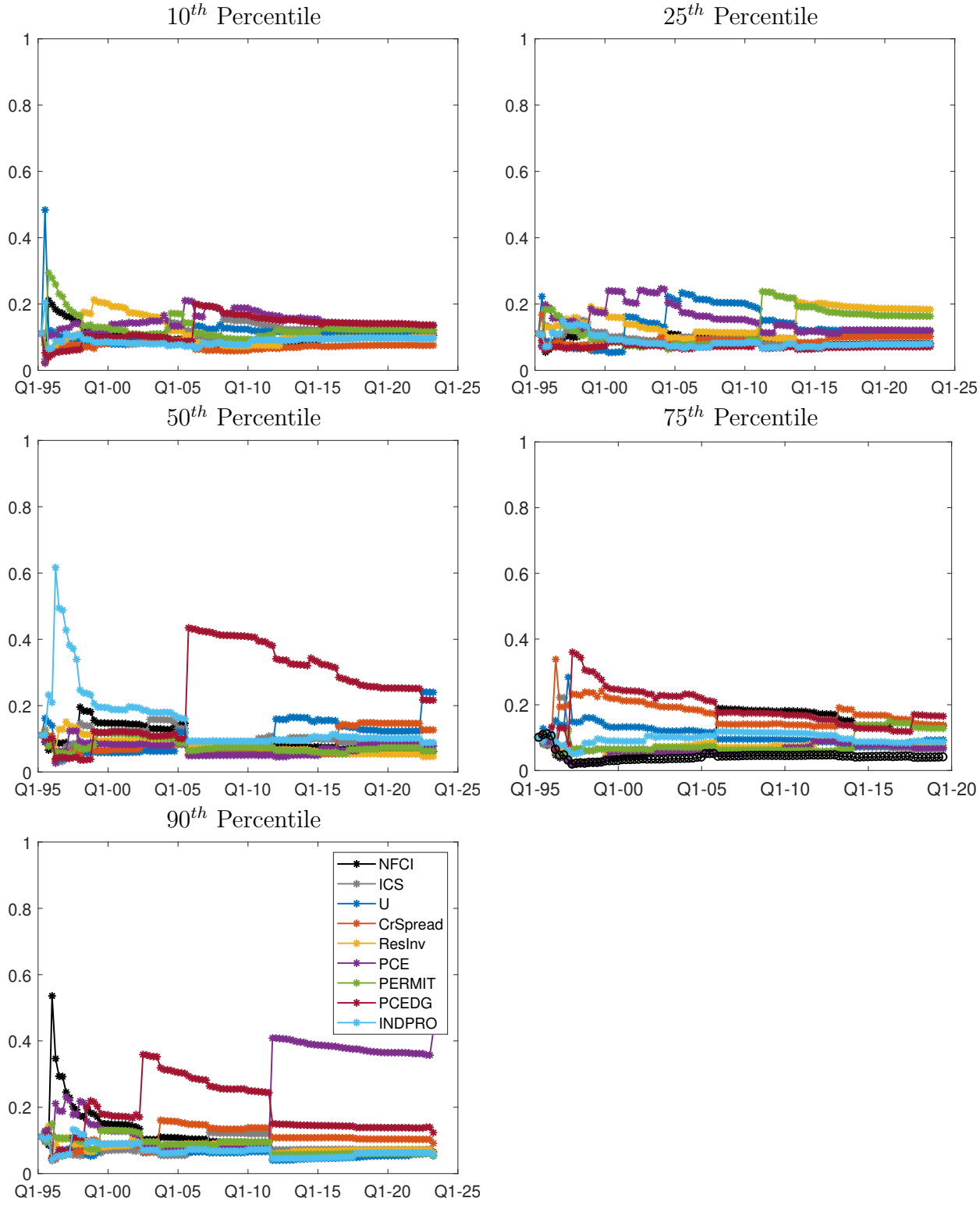


Figure A.10: Calibration of quantile-combined density forecast. Empirical CDF for PITs with the empirical 5% critical values calculated using Bootstrap following Rossi and Sekhposyan (2019). Sample up to 2023Q3, including the COVID-19 pandemic.

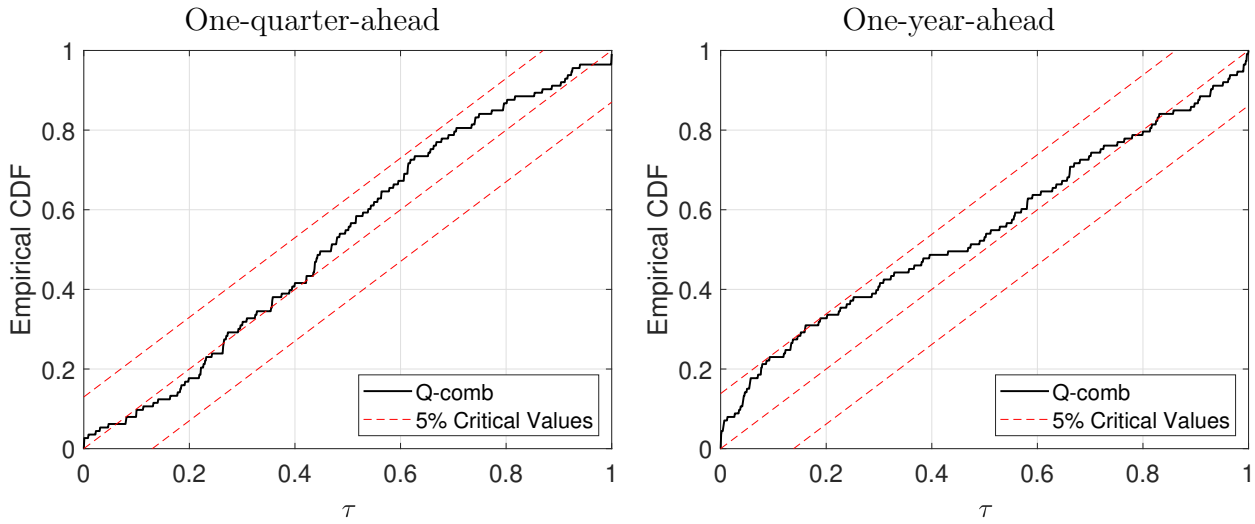


Table A.10: Calibration tests statistics for quantile-combined forecasts - sample up to 2023Q3, including pandemic.

one-quarter-ahead combined density forecasts

Calibration Tests	Statistic	Bootst. Critical Values at		
		10% s.l.	5% s.l.	1% s.l.
Kolmogorov–Smirnov	1.143	1.384	1.485	1.844
Cramér–von Mises	0.275	0.398	0.497	0.738
	p-values			
Knüppel NW	0.887			

one-year-ahead combined density forecasts

Calibration Tests	Statistic	Bootst. Critical Values at		
		10% s.l.	5% s.l.	1% s.l.
Kolmogorov–Smirnov	1.581	1.473	1.756	2.179
Cramér–von Mises	0.633	0.665	0.854	2.042
	p-values			
Knüppel NW	0.702			

Note: The null hypothesis of calibration is rejected for Kolmogorov–Smirnov and Cramér–von Mises if the test statistic is greater than the bootstrapped critical values following Rossi and Sekhposyan (2019); for Knüppel (2015) test rejects null hypothesis of calibration if the p-value displayed here is lower than significance level.

A.5 Additional results when using alternative GDP vintages

A.5.1 First Release of GDP

Figure A.11: Quantile combination weights over time for one-quarter-ahead forecasts for all $K = 9$ forecasting models - first release of GDP.

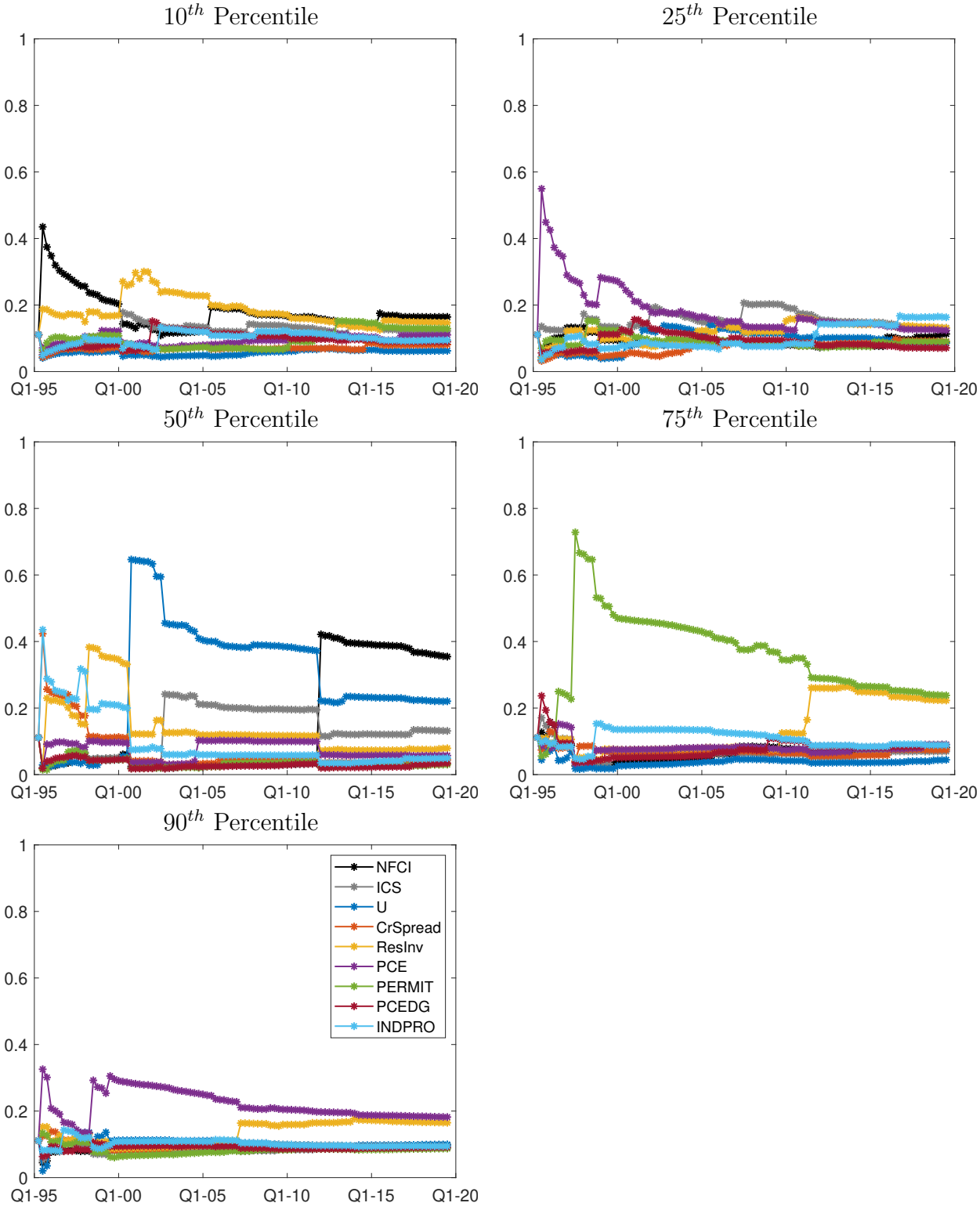


Figure A.12: Quantile combination weights over time for one-year-ahead forecasts for all $K = 9$ forecasting models - first release of GDP.

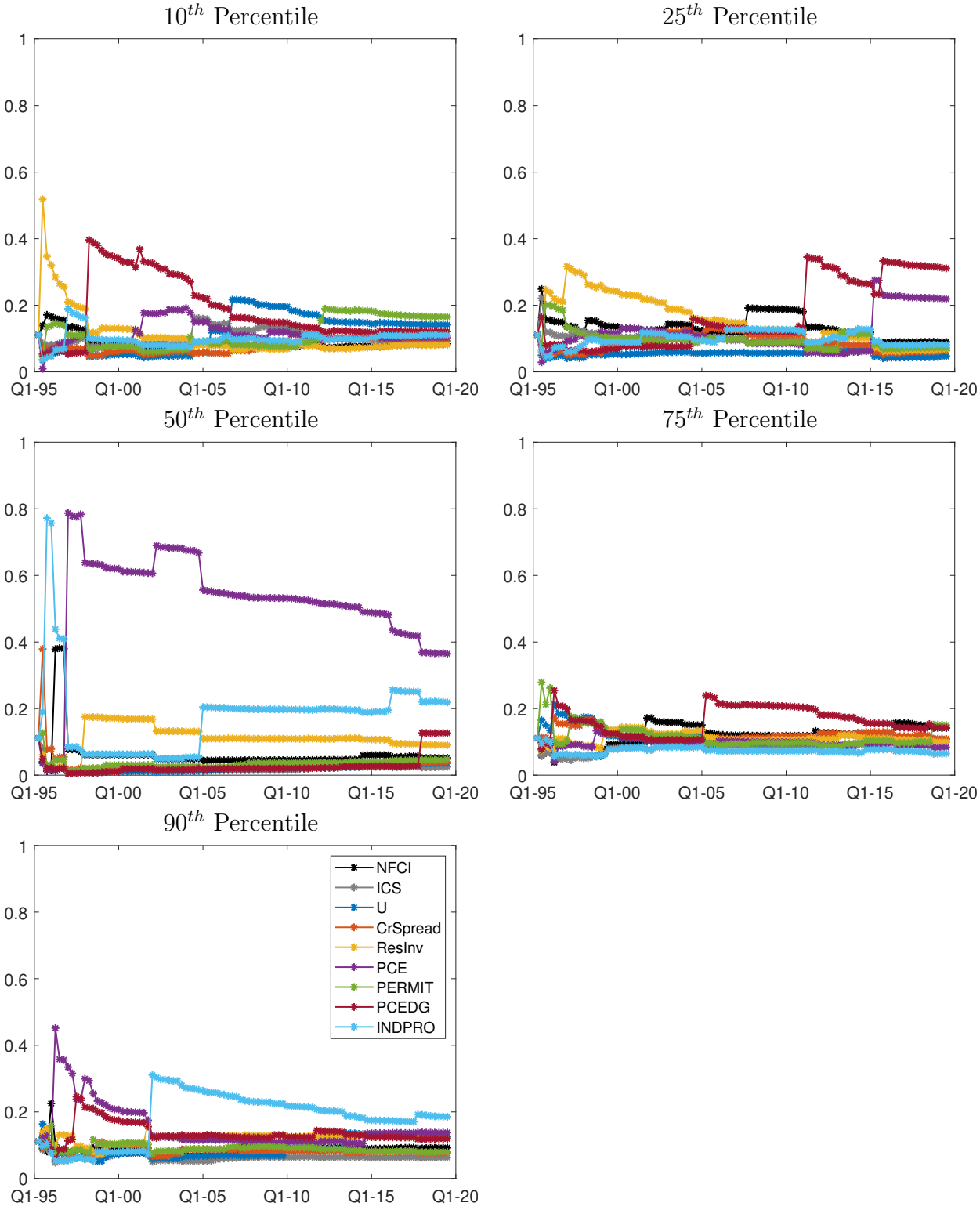


Figure A.13: Cumulative CRPS of the alternative combination approaches relative to quantile combination for one-quarter and one-year ahead forecasts - first release of GDP.

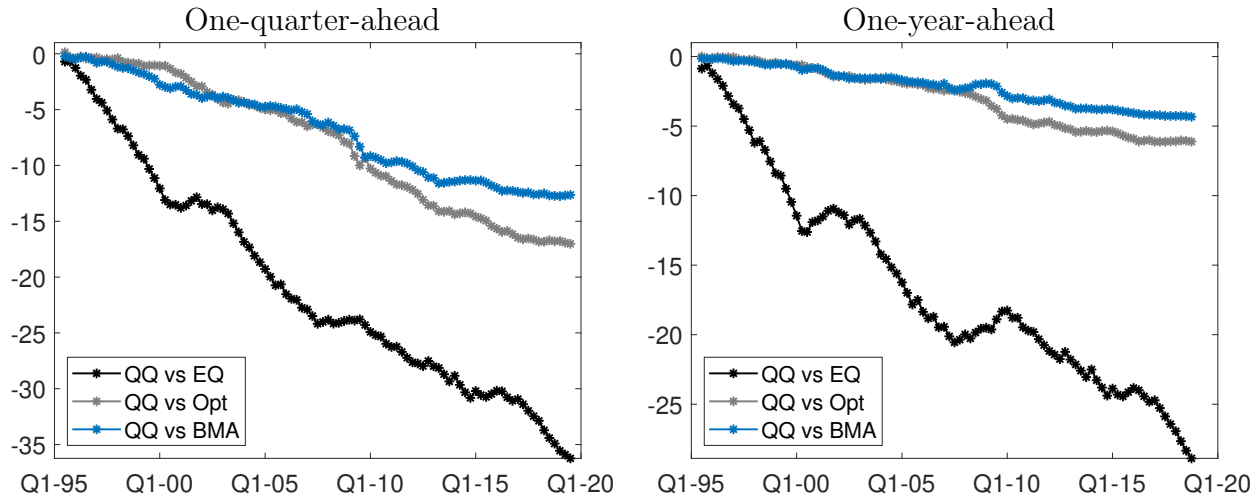


Figure A.14: Calibration of quantile-combined density forecast. Empirical CDF for PITs with the empirical 5% critical values calculated using Bootstrap following Rossi and Sekhposyan (2019) - first release of GDP.

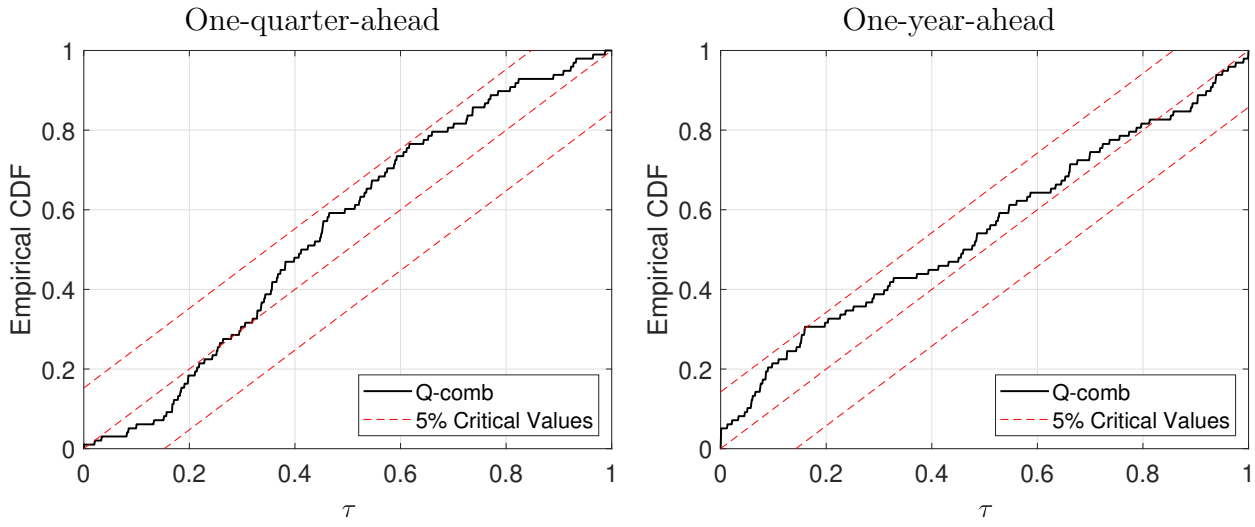


Table A.11: Calibration tests statistics for quantile-combined forecasts - first release of GDP.

one-quarter-ahead combined density forecasts

Calibration Tests	Statistic	Bootst. Critical Values at		
		10% s.l.	5% s.l.	1% s.l.
Kolmogorov–Smirnov	1.468	1.505	1.737	2.136
Cramér–von Mises	0.627	0.580	0.806	1.324
	p-values			
Knüppel NW	0.863			

one-year-ahead combined density forecasts

Calibration Tests	Statistic	Bootst. Critical Values at		
		10% s.l.	5% s.l.	1% s.l.
Kolmogorov–Smirnov	1.446	1.413	1.572	2.076
Cramér–von Mises	0.403	0.553	0.736	0.970
	p-values			
Knüppel NW	0.735			

Note: The null hypothesis of calibration is rejected for Kolmogorov–Smirnov and Cramér–von Mises if the test statistic is greater than the bootstrapped critical values following Rossi and Sekhposyan (2019); for Knüppel (2015) test rejects null hypothesis of calibration if the p-value displayed here is lower than significance level.

A.5.2 Latest available release of GDP

Figure A.15: Quantile combination weights over time for one-quarter-ahead forecasts for all $K = 9$ forecasting models - latest available release of GDP.

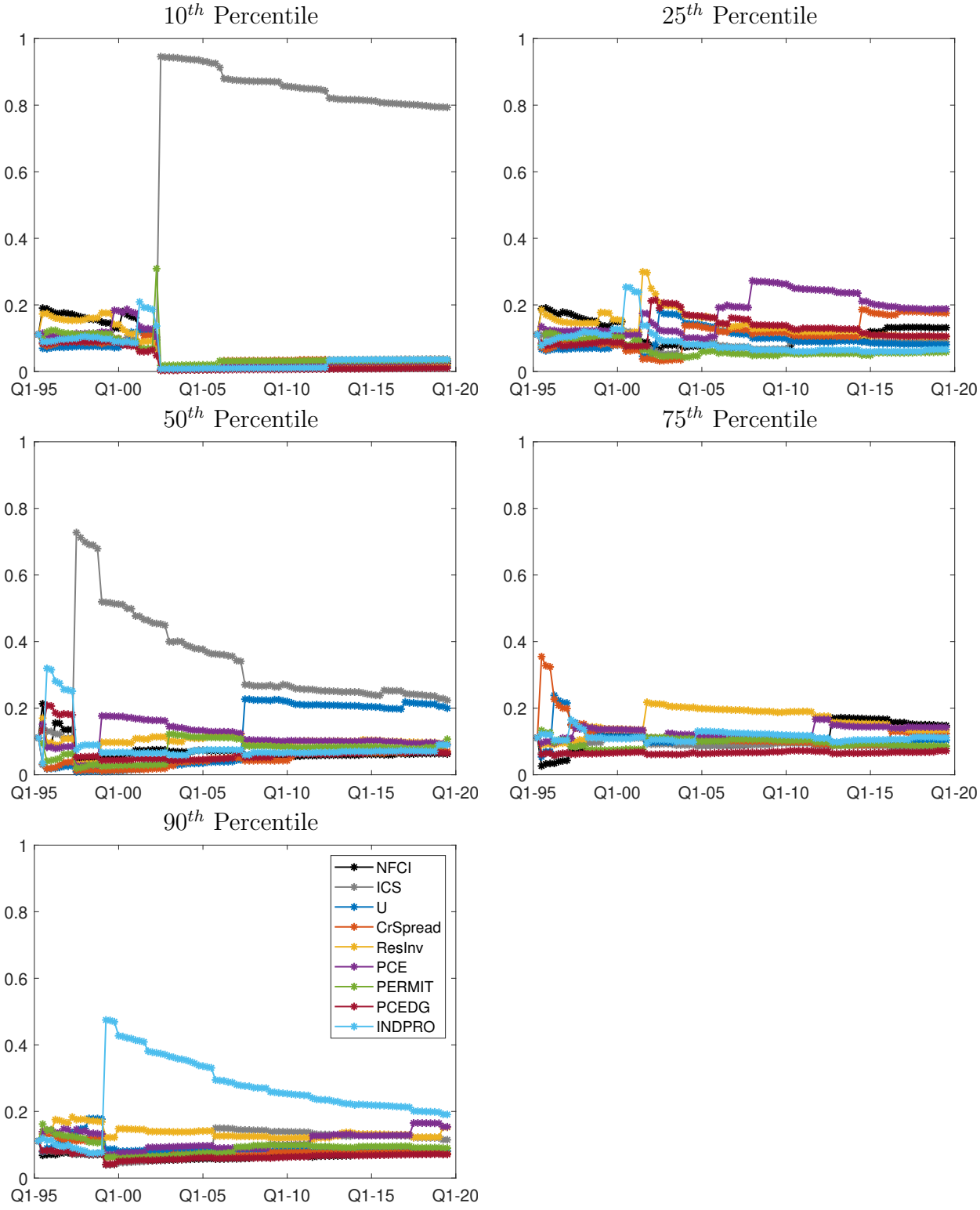


Figure A.16: Quantile combination weights over time for one-year-ahead forecasts for all $K = 9$ forecasting models - latest available release of GDP.

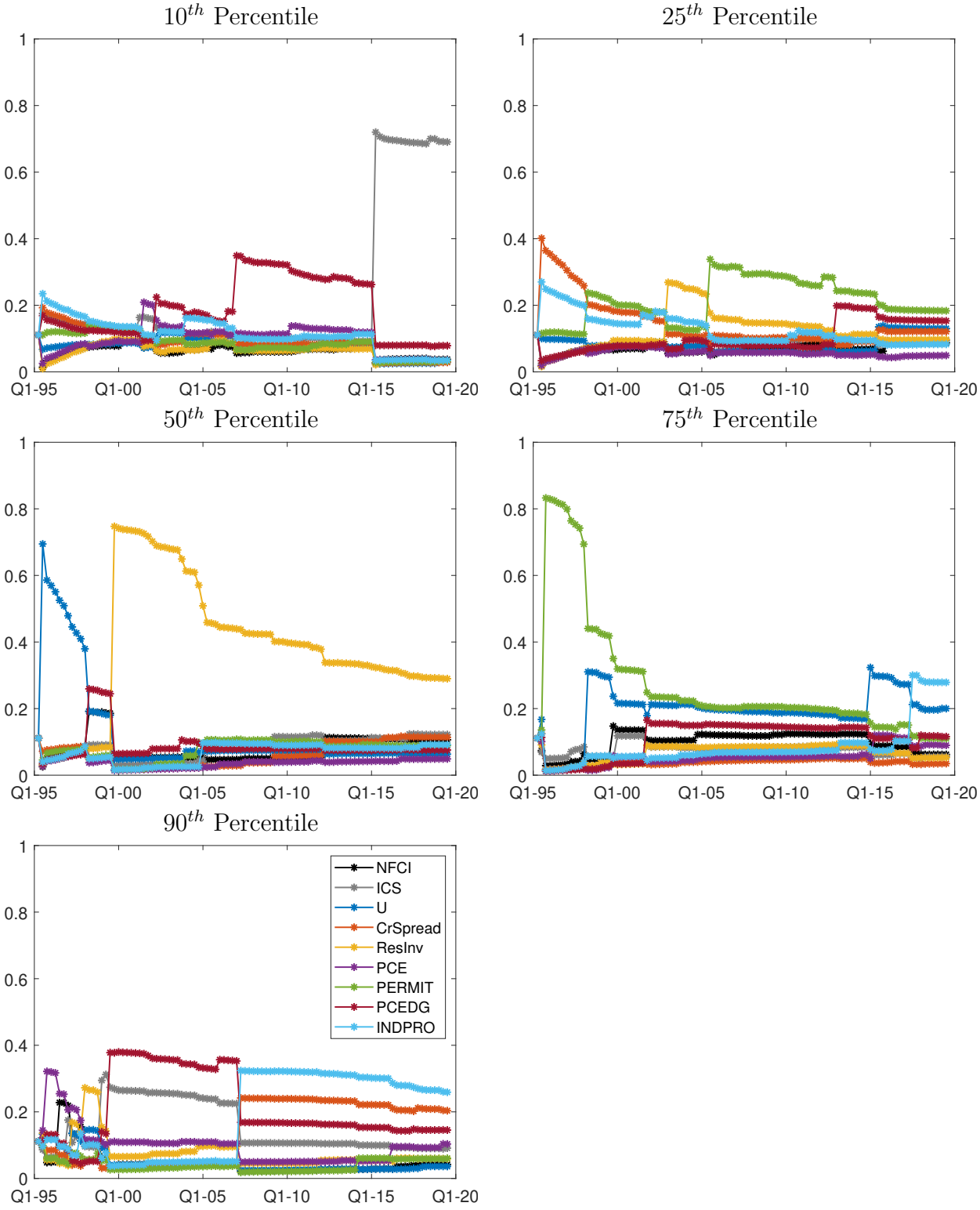


Figure A.17: Cumulative CRPS of the alternative combination approaches relative to quantile combination for one-quarter and one-year ahead forecasts - latest available release of GDP.

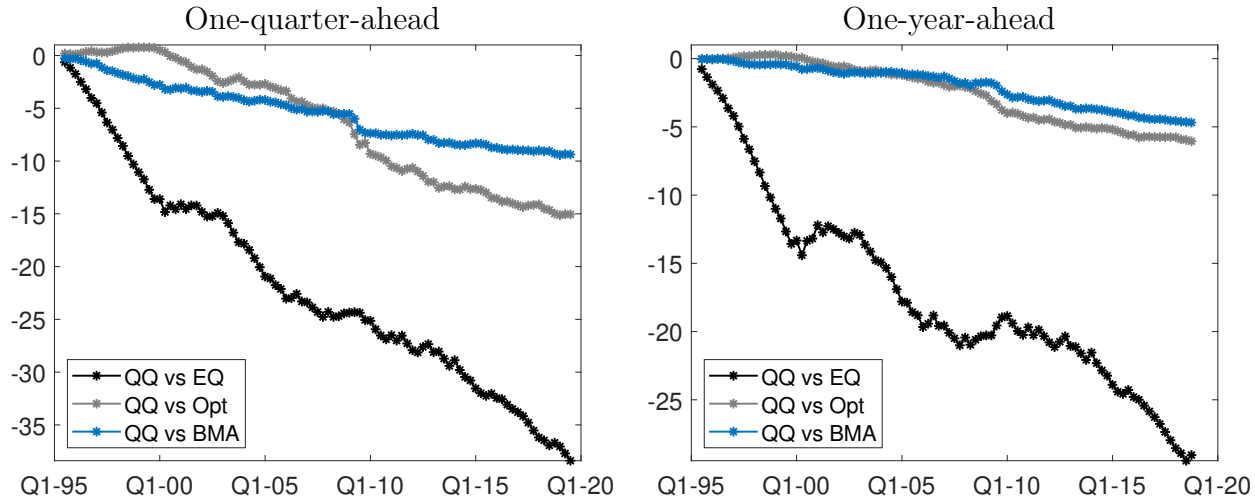


Figure A.18: Calibration of quantile-combined density forecast. Empirical CDF for PITs with the empirical 5% critical values calculated using Bootstrap following Rossi and Sekhposyan (2019) - latest available release of GDP.

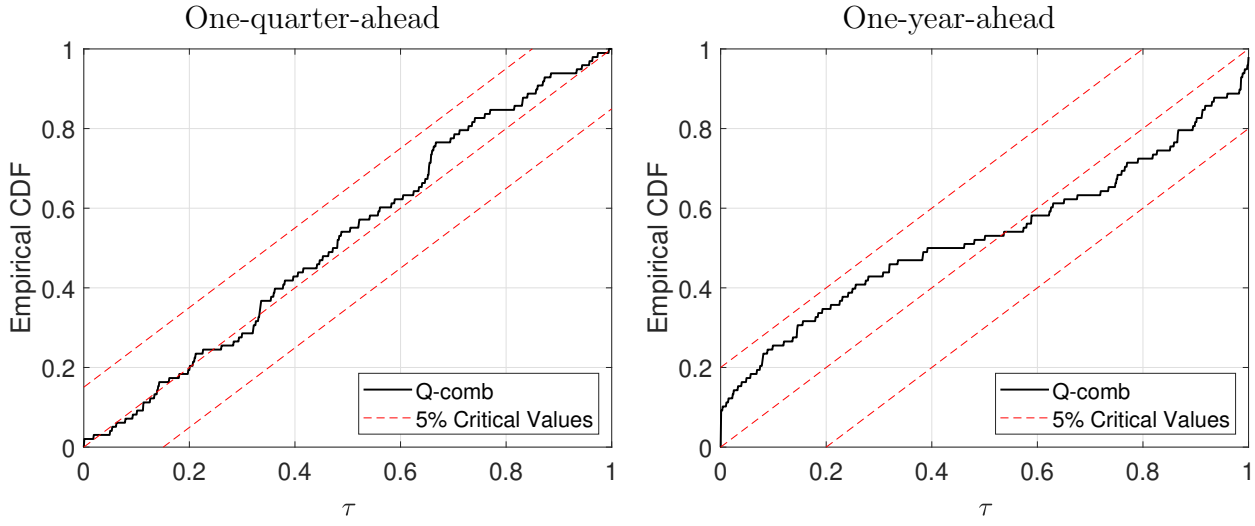


Table A.12: Calibration tests statistics for quantile-combined forecasts - latest release.

one-quarter-ahead combined density forecasts

Calibration Tests	Statistic	Bootst. Critical Values at		
		10% s.l.	5% s.l.	1% s.l.
Kolmogorov–Smirnov	0.973	1.491	1.705	2.295
Cramér–von Mises	0.135	0.645	0.824	1.747
	p-values			
Knüppel NW	0.996			

one-year-ahead combined density forecasts

Calibration Tests	Statistic	Bootst. Critical Values at		
		10% s.l.	5% s.l.	1% s.l.
Kolmogorov–Smirnov	1.587	1.976	2.170	2.539
Cramér–von Mises	0.923	1.426	1.801	2.763
	p-values			
Knüppel NW	0.784			

Note: The null hypothesis of calibration is rejected for Kolmogorov–Smirnov and Cramér–von Mises if the test statistic is greater than the bootstrapped critical values following Rossi and Sekhposyan (2019); for Knüppel (2015) test rejects null hypothesis of calibration if the p-value displayed here is lower than significance level.