Taylor & Francis
Taylor & Francis Group

Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ubes20

Testing Quantile Forecast Optimality

Jack Fosten, Daniel Gutknecht & Marc-Oliver Pohle

To cite this article: Jack Fosten, Daniel Gutknecht & Marc-Oliver Pohle (12 Mar 2024): Testing Quantile Forecast Optimality, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2024.2316091

To link to this article: https://doi.org/10.1080/07350015.2024.2316091

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2024, VOL. 00, NO. 0, 1–12
https://doi.org/10.1080/07350015.2024.2316091

Taylor & Francis
Taylor & Francis Group

Testing Quantile Forecast Optimality

Jack Fosten[a], Daniel Gutknecht[b], and Marc-Oliver Pohle[c]

[a]King's Business School, King's College London, London, UK; [b]Faculty of Economics and Business, Goethe University Frankfurt, Frankfurt am Main, Germany; [c]Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

ABSTRACT
Quantile forecasts made across multiple horizons have become an important output of many financial institutions, central banks and international organizations. This article proposes misspecification tests for such quantile forecasts that assess optimality over a set of multiple forecast horizons and/or quantiles. The tests build on multiple Mincer-Zarnowitz quantile regressions cast in a moment equality framework. Our main test is for the null hypothesis of autocalibration, a concept which assesses optimality with respect to the information contained in the forecasts themselves. We provide an extension that allows to test for optimality with respect to larger information sets and a multivariate extension. Importantly, our tests do not just inform about general violations of optimality, but may also provide useful insights into specific forms of sub-optimality. A simulation study investigates the finite sample performance of our tests, and two empirical applications to financial returns and U.S. macroeconomic series illustrate that our tests can yield interesting insights into quantile forecast sub-optimality and its causes.

ARTICLE HISTORY
Received February 2023
Accepted December 2023

KEYWORDS
Forecast evaluation; Forecast rationality; Multiple horizons; Quantile regression

1. Introduction

Economic and financial forecasters have become increasingly interested in making quantile predictions, often across different quantile levels and at multiple horizons into the future. In financial markets, for instance, such multi-step quantile predictions are produced due to the 10-day value-at-risk (VaR) requirements of the Basel Committee on Banking Supervision.[1] In the growth-at-risk (GaR) literature on the other hand, Adrian, Boyarchenko, and Giannone (2019) propose quantile models to predict downside risks to real gross domestic product (GDP) growth at horizons ranging from one quarter ahead to one year ahead. These methods are now widely implemented in academic research (Plagborg-Møller et al. 2020; Brownlees and Souza 2021) and in international institutions like the IMF (Prasad et al. 2019), and are typically applied across various quantile levels. This trend for multi-horizon quantile forecasts has also developed into a growing literature in nowcasting GaR that typically uses several intra-period nowcast horizons (e.g., Carriero, Clark, and Marcellino 2020; Antolin Diaz, Drechsel, and Petrella 2021; Ferrara, Mogliani, and Sahuc 2021). Finally, it is common for central banks, such as the Bank of England, to produce fan charts of key economic variables such as GDP growth, unemployment or the Consumer Price Index (CPI) inflation rate across several quantile levels and horizons.

However, despite the expansion in empirical and methodological research, there is currently very little statistical guidance for assessing whether a set of multi-step ahead, multi-quantile forecasts are consistent with respect to the outcomes observed. This consistency is often referred to as "optimality," "rationality," or "calibration" in the literature, with 'full optimality' referring to optimality relative to the information set known to the forecaster, while a weaker form of optimality known as 'autocalibration' is defined with respect to the information contained in the forecasts themselves (see Gneiting and Ranjan 2013; Tsyplakov 2013). This article aims to fill this gap in the literature by proposing various (out-of-sample) optimality tests for quantile forecasts that can accommodate predictions either derived from known econometric forecasting models, or from external sources like institutional or professional forecasters. Specifically, we develop tests that assess optimality of quantile forecasts over multiple forecast horizons *and* multiple quantiles simultaneously.

The main test of this article is a joint test of autocalibration for quantile forecasts obtained across different horizons and quantile levels. The test is based on a series of quantile Mincer-Zarnowitz (MZ) regressions (see Gaglianone et al. 2011) across all quantile levels and horizons, which are in turn used to construct a test statistic for the null hypothesis of autocalibration across horizons and quantiles using a set of moment equalities (e.g., Romano and Shaikh 2010; Andrews and Soares 2010). We suggest a block bootstrap procedure to obtain critical values for the test. The bootstrap is simple to implement and avoids the need to estimate a large variance-covariance matrix that would be required in a more standard Wald-type test. We establish the first-order asymptotic validity of these bootstrap critical values.

CONTACT Daniel Gutknecht Gutknecht@wiwi.uni-frankfurt.de Faculty of Economics and Business, Goethe University Frankfurt, Frankfurt am Main, Germany.

[1] See for instance: *https://www.bis.org/publ/bcbs148.pdf* [Last Accessed: 18/12/20]

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/UBES*.

The test of autocalibration based on MZ regressions can provide valuable information to forecasters. In particular, failure to reject the null hypothesis of autocalibration suggests that the forecaster may proceed to use the forecasts as they are without the need to "re-calibrate" them. On the other hand, if the null hypothesis is rejected, the test hints at directions for improvement of the forecasts. That is, it informs the forecaster about the horizons, quantiles or horizon-quantile combinations that contributed strongest to the rejection of the null, and thus an improvement of the forecasts is warranted. In addition, the estimated MZ regression can be used to infer about the nature of the deviations from autocalibration, when the forecasts are plotted alongside the realizations. The estimated MZ coefficients may also be used to perform a re-calibration of the original forecasts, as has been suggested in the case of mean forecasts in the recent work of Clements (2022).

We provide two extensions of this test for autocalibration. The first extension allows for additional predictors in the MZ regressions, which we call the augmented quantile Mincer-Zarnowitz test. This test is operationally similar to the first test, but may provide richer information to the forecaster. It tests a stronger form of optimality relative to a larger information set than autocalibration. If autocalibration is not rejected, but the null hypothesis of the augmented test is rejected, it indicates that the additional variables used in the MZ regression carry additional informational content which should be used in making the forecasts themselves. The second extension allows to test optimality for multiple time series variables and not just for a single variable. Testing multiple time series variables simultaneously may be useful in cases where we are interested in testing whether one type of model delivers optimal forecasts for multiple macroeconomic variables, or across different financial asset returns, for instance different companies from the same sector.

Finally, as a separate contribution, we also outline in the supplementary material a test for monotonically nondecreasing expected quantile loss as the forecast horizon increases. This extends the result of Patton and Timmermann (2012) to the quantile case whereas they focussed on the mean squared forecast error (MSFE) case for optimal multi-horizon mean forecasts. The test makes use of empirical moment inequalities using the Generalized Moment Selection (GMS) procedure of Andrews and Soares (2010). This test can also be seen as complementary to monotonicity tests used in the nowcasting literature for the MSFE of mean nowcasts (see Fosten and Gutknecht 2020, and references therein).

We provide two empirical applications of our methodology. The first one applies the basic MZ test to classical VaR forecasts for S&P 500 returns constructed from a GARCH(1,1) model via the GARCH bootstrap (Pascual, Romo, and Ruiz 2006). We test jointly over the quantile levels 0.01, 0.025, and 0.05 and horizons from 1 to 10 trading days. Autocalibration is rejected overall and the miscalibration of the forecasts gets stronger for larger forecast horizons and more extreme quantiles. Furthermore, a clear pattern emerges over all quantiles and horizons regarding the conditional quantile bias: the VaR forecasts tend to underestimate risk in calmer times, but overestimate it in more stressful periods.

The second empirical study applies the test in the spirit of the emerging GaR literature, where we focus on the extensions of our test using the augmented MZ test and the test with multiple time series. We expand on the work of Adrian, Boyarchenko, and Giannone (2019) to formally investigate the performance of simple quantile regression models using financial conditions indicators in predicting a range of U.S. macroeconomic series. Interestingly, we find that the forecasts across four different series and a range of quantile levels and horizon are sub-optimal in that they are not autocalibrated. However, further analysis of the results shows that this sub-optimality is present only in inflation-type series and not in real series like industrial production and employment growth. We also find poorer calibration at the most extreme quantile under consideration.

In relation to the existing literature, this article extends the work on quantile forecast optimality or, in other words, absolute evaluation of quantile forecasts. The focus of this literature has been on single-horizon prediction at a single quantile, which mainly stems from the extensive body of research on backtesting VaR, such as Christoffersen (1998), Engle and Manganelli (2004), Escanciano and Olmo (2010, 2011), Gaglianone et al. (2011), and Nolde and Ziegel (2017). Our work also complements the literature on testing the relative forecast performance of conditional quantile models such as Giacomini and Komunjer (2005), Manzan (2015) or more recently Corradi, Fosten, and Gutknecht (2023). Finally, as our focus lies on testing for optimality across horizons, the article also relates to Quaedvlieg (2021), who emphasized the importance of multi-horizon forecast evaluation to avoid multiple testing issues in the context of relative evaluation of mean forecasts. The only work on multi-horizon optimality testing we are aware of is Patton and Timmermann (2012), who consider the case of mean forecasts as well and discuss several implications of optimality specific to the multi-horizon context and how to construct tests for them, most notably the monotonicity of expected loss over horizons, which we extend to the quantile case in the supplementary material.

The rest of the article is organized as follows. Section 2 lays out the notion of quantile forecast optimality that will provide the foundation of our tests. Section 3 then introduces the test for autocalibration via MZ regression, along with the bootstrap methodology and theory. Section 4 extends the test to the augmented MZ test and the test for multiple variables, while Section 5 gives the two empirical applications of our methods. Finally, Section 6 concludes the article.

The supplementary material contains results from a Monte Carlo study (Section S1), where we assess the finite sample properties of the MZ and augmented MZ tests across various sample sizes and bootstrap block lengths. The supplement also contains the proofs for the theoretical results (Sections S2 and S3) along with the monotonicity test (Section S4). Sections S5 and S6 provide additional empirical results and graphs for the VaR and the GaR application, respectively. Finally, all tests of the article are provided as `R` functions in the `R` package `quantoptimR` available at *https://github.com/MarcPohle/quantoptimR*.

2. Quantile Forecast Optimality

Consider a multivariate stochastic process $\{\mathbf{V}_t\}_{t\in\mathbb{Z}}$, where $\mathbf{V}_t$ is a random vector which contains a response variable of interest y_t and other observable predictors. We denote the forecaster's information set at time t by $\mathcal{F}_t = \sigma(\mathbf{V}_s; s \leq t)$, where $\sigma(.)$ denotes the σ-algebra generated by a set of random variables. Assuming a continuous outcome y_t with strictly positive density everywhere for the rest of the article, our target functional is the conditional τ-quantile of y_t given $\mathcal{F}_{t-h}$:

$$q_t\left(\tau|\mathcal{F}_{t-h}\right) = F^{-1}_{y_t|\mathcal{F}_{t-h}}(\tau),$$

where $F_{y_t|\mathcal{F}_{t-h}}(\cdot)$ is the cumulative distribution function of y_t conditional on $\mathcal{F}_{t-h}$. We denote an h-step ahead forecast at time $t-h$ for this τ-quantile $q_t\left(\tau|\mathcal{F}_{t-h}\right)$ by $\widehat{y}_{\tau,t,h}$, and assume that we observe these forecasts $\widehat{y}_{\tau,t,h}$ for each target period t at multiple horizons, $h \in \mathcal{H} = \{1,\dots,H\}$, and multiple quantile levels, $\tau \in \mathcal{T} = \{\tau_1,\dots,\tau_K\} \subset [0+\varepsilon, 1-\varepsilon]$ with $\varepsilon > 0$, for some finite integers H and K, respectively. That is, at each time point t we have a matrix of forecasts, $\left(\widehat{y}_{\tau,t,h}\right)_{\tau=\tau_1,\dots,\tau_K, h=1,\dots,H}$. In addition, throughout the article, we will assume strict stationarity of $\{\mathbf{V}_t\}_{t\in\mathbb{Z}}$ and finite first moments of the forecasts $\widehat{y}_{\tau,t,h}$ and y_t itself, see Assumptions A1 and A2 in Section 3.

Since our focus lies on the evaluation of quantile forecasts, the loss function used for evaluation in this context is the "tick" or "check" loss which is well-known from quantile regression. This is written as $L_\tau\left(y_{t+h} - \widehat{y}_{\tau,t,h}\right) = \rho_\tau\left(y_{t+h} - \widehat{y}_{\tau,t,h}\right)$, where $\rho_\tau(u) = u\,(\tau - 1\{u < 0\})$ and where $1\{.\}$ denotes the indicator function giving a value of one when the expression is true and zero otherwise.

While relative forecast evaluation deals with comparing different forecasting methods or models, mainly by ranking them via their expected loss, the subject of this article is absolute forecast evaluation across different quantile levels and/or forecasting horizons, in other words the assessment of *a particular* forecasting model or method in terms of absolute evaluation criteria for multiple quantile levels and horizons. These evaluation criteria are usually different forms of optimality (or "rationality"/"calibration"). We start by defining and discussing various forms of quantile forecast optimality before showing how to operationalize the latter for testing.

Definition 1 (Optimality). An h-step ahead forecast $\widehat{y}^*_{\tau,t,h|\mathcal{I}_{t-h}}$ for the τ-quantile is optimal relative to an information set $\mathcal{I}_{t-h} \subset \mathcal{F}_{t-h}$ if:

$$\widehat{y}^*_{\tau,t,h|\mathcal{I}_{t-h}} = \arg\min_{\widehat{y}_{\tau,t,h}} \mathrm{E}\left[L_\tau\left(y_t - \widehat{y}_{\tau,t,h}\right)|\mathcal{I}_{t-h}\right].$$

We simply call it optimal and denote it by $\widehat{y}^*_{\tau,t,h}$ if $\mathcal{I}_{t-h} = \mathcal{F}_{t-h}$, that is if it is optimal relative to the full information set: $\widehat{y}^*_{\tau,t,h} \equiv \widehat{y}^*_{\tau,t,h|\mathcal{F}_{t-h}}$.

Analogous to the case of mean forecasts (Granger 1969), an optimal quantile forecast relative to an information set can alternatively be characterized as being equal to the respective conditional quantile provided the information set is sufficiently large and includes the forecasts themselves. Specifically, since "tick" loss is a strictly consistent scoring function for the corresponding quantile (see Definition 1 and Proposition 1 in Gneiting 2011), it holds that an h-step ahead forecast $\widehat{y}_{\tau,t,h}$ for the τ-quantile is optimal relative to any information (sub-)set $\mathcal{I}_{t-h}$ satisfying $\sigma\left(\widehat{y}_{\tau,t,h}\right) \subset \mathcal{I}_{t-h} \subset \mathcal{F}_{t-h}$, where $\sigma\left(\widehat{y}_{\tau,t,h}\right)$ denotes the sigma algebra spanned by the forecast itself, if and only if:

$$\widehat{y}_{\tau,t,h} = q_t\left(\tau|\mathcal{I}_{t-h}\right).[2] \qquad (1)$$

While interest often lies in testing the null hypothesis of (full) optimality relative to the information set $\mathcal{F}_{t-h}$, which amounts to testing if the forecast, $\widehat{y}_{\tau,t,h}$, is equal to its target, $q_t(\tau|\mathcal{F}_{t-h})$, the possibly large and generally unknown information set $\mathcal{F}_{t-h}$ usually makes direct tests of this hypothesis difficult in practice. Thus, we next discuss weaker forms of optimality that will form the basis of our test(s) in Sections 3 and 4. In fact, in Section S2 of the supplement (see Lemma S.1) we show formally that these weaker forms of optimality may always be viewed as a direct implication of optimality with respect to the "full" information set $\mathcal{F}_{t-h}$. That is, any h-step ahead forecast optimal with respect to the full information set $\mathcal{F}_{t-h}$, is also optimal relative to any 'smaller' information (sub-)set $\mathcal{I}_t \subset \mathcal{F}_t$.

A special case of this "weaker" form of optimality is optimality with respect to the information contained in the forecast itself, $\sigma\left(\widehat{y}_{\tau,t,h}\right)$, or autocalibration, a term first coined by Tsyplakov (2013) and Gneiting and Ranjan (2013) in the context of probabilistic forecasts.

Definition 2 (Autocalibration). An h-step ahead forecast $\widehat{y}_{\tau,t,h}$ for the τ-quantile is autocalibrated if it holds that: $\widehat{y}_{\tau,t,h} = q_t\left(\tau|\sigma(\widehat{y}_{\tau,t,h})\right)$.

On the one hand, autocalibration may be regarded as a direct implication of full optimality that is particularly suitable for testing as it only relies upon the forecasts themselves and does not require any assumptions on the information set $\mathcal{F}_{t-h}$ or a selection of variables from it. On the other hand, however, autocalibration may also be viewed as a criterion for absolute forecast evaluation in its own right for several reasons. First, the concept has a clear interpretation since a forecast user provided with autocalibrated forecasts should use them as they are and not transform or "recalibrate" them. Second, only involving forecasts and observations and no information set that depends on other quantities, it comes closest to the idea of forecast calibration as a concept of consistency between forecasts and observations (see Gneiting, Balabdaoui, and Raftery 2007). Third, autocalibration might often be a more reasonable criterion to demand from forecasts than full optimality, which is a often hard to fulfill in practice. Finally, the Murphy decomposition of expected loss (Pohle 2020) shows that autocalibration is a fundamental property of forecasts in that expected loss is driven by only two forces: deviations from autocalibration and the information content of the forecasts. The next section will outline how to test autocalibration, viewed either as an implication of full optimality or as a forecast property in its own right, across multiple quantile levels and horizons simultaneously.

[2] Note that this result continues to hold for any quantile loss from the class of generalized piecewise linear loss functions, or in fact for any consistent scoring function if the τ-quantile is substituted for the corresponding statistical functional for which this scoring function is consistent.

3. Quantile Mincer-Zarnowitz Test

3.1. Null Hypothesis and Quantile Mincer-Zarnowitz Regressions

While autocalibration testing has a long tradition in econometrics through the use of Mincer-Zarnowitz regressions for mean forecasts (Mincer and Zarnowitz 1969), the latter may also be used directly for the case of quantiles (see Gaglianone et al. 2011). Definition 2 in fact suggests that a natural test for autocalibration of an h-step ahead forecast for the τ-level quantile may be based on checking whether, for a given sample of outcomes and quantile forecasts at level τ_k and horizon h, it holds that:

$$q_t\left(\tau|\widehat{y}_{\tau,t,h}\right) = \alpha_h^{\dagger}(\tau_k) + \widehat{y}_{\tau_k,t,h}\beta_h^{\dagger}(\tau_k) = \widehat{y}_{\tau_k,t,h}$$

almost surely. More generally, since our goal is to test for autocalibration over multiple forecast horizons and quantile levels jointly, we specify such a linear quantile regression model for every horizon $h \in \mathcal{H}$ and $\tau_k \in \mathcal{T}$ as follows:

$$\begin{aligned} y_t &= \alpha_h^{\dagger}(\tau_k) + \widehat{y}_{\tau_k,t,h}\beta_h^{\dagger}(\tau_k) + \varepsilon_{t,h}(\tau_k) \\ &= \mathbf{X}'_{\tau_k,t,h}\boldsymbol{\beta}_h^{\dagger}(\tau_k) + \varepsilon_{t,h}(\tau_k), \end{aligned} \tag{2}$$

where $\mathbf{X}_{\tau_k,t,h} = (1, \widehat{y}_{\tau_k,t,h})'$ and $\boldsymbol{\beta}_h^{\dagger}(\tau_k) = (\alpha_h^{\dagger}(\tau_k), \beta_h^{\dagger}(\tau_k))'$. Here, the population coefficient vector $\boldsymbol{\beta}_h^{\dagger}(\tau_k) = (\alpha_h^{\dagger}(\tau_k), \beta_h^{\dagger}(\tau_k))'$ of this linear quantile regression model is defined as

$$\boldsymbol{\beta}_h^{\dagger}(\tau_k) = \arg\min_{\boldsymbol{b}\in\mathcal{B}} \mathrm{E}\left[\rho_{\tau_k}\left(y_t - \mathbf{X}'_{\tau_k,t,h}\boldsymbol{b}\right)\right], \tag{3}$$

where $\mathcal{B}$ denotes the parameter space satisfying conditions set out in Assumption A3. The composite null hypothesis is given by:

$$H_0^{\mathrm{MZ}} : \{\alpha_h^{\dagger}(\tau_k) = 0\} \cap \{\beta_h^{\dagger}(\tau_k) = 1\} \tag{4}$$

for all $h \in \mathcal{H}$ and $\tau_k \in \mathcal{T}$ versus $H_1^{\mathrm{MZ}} : \{\alpha_h^{\dagger}(\tau_k) \neq 0\}$ and/or $\{\beta_h^{\dagger}(\tau_k) \neq 1\}$ for at least some $h \in \mathcal{H}$ and $\tau_k \in \mathcal{T}$. Testing the null hypothesis in (4) not only yields a multi-horizon, multi-quantile test of autocalibration, but also provides us with an idea about possible deviations from the null. In particular, examining the contributions of single horizons, quantiles or horizon-quantile combinations to the overall test statistic, which will be introduced in Section 3.2, may also be informative about deviations from autocalibration. Moreover, the empirical counterpart of $q_t\left(\tau|\widehat{y}_{\tau,t,h}\right) = \alpha_h^{\dagger}(\tau_k) + \widehat{y}_{\tau_k,t,h}\beta_h^{\dagger}(\tau_k)$ may be interpreted as autocalibrated forecasts such that, for a specific value of the forecast $\widehat{y}_{\tau_k,t,h}$, the deviations between the regression line (or recalibrated forecast) and the forecasts themselves, $\widehat{y}_{\tau_k,t,h} - q_t\left(\tau|\widehat{y}_{\tau,t,h}\right)$ can be interpreted as the quantile version of a conditional bias. The direction and size of this conditional quantile bias informs us about deficiencies of forecasts in certain situations, a point that we will come back to and illustrate in the applications in Section 5.

3.2. Test Statistic and Bootstrap

In what follows, assume that we observe an evaluation sample of size P, in other words a scalar-valued time series of observations starting at some point in time $R+1$, $\{y_t\}_{t=R+1}^{T}$, and a matrix-valued time series of forecasts, $\left\{\left(\widehat{y}_{\tau,t,h}\right)_{\tau=\tau_1,\dots,\tau_K,h=1,\dots,H}\right\}_{t=R+1}^{T}$. Moreover, we may also observe a vector of additional variables $\mathbf{Z}_{t-h}$ from the forecaster's information set $\mathcal{F}_{t-h}$. We will write this additional sample of vector-valued time series as $\{\mathbf{Z}_t\}_{t=R+1-H}^{T-1}$.

Using the taxonomy of Giacomini and Rossi (2010), forecasts $\widehat{y}_{\tau_k,t,h}$ may stem either from "forecasting methods" or from "forecasting models". In the former case, we are typically without knowledge about the underlying model such as with forecasts from the Survey of Professional Forecasters (SPF), or may use forecasts that depend on parameters estimated in-sample using so-called limited-memory estimators based on a finite rolling estimation window. In the case of "forecasting models" on the other hand, we need to account for the contribution of estimation uncertainty to the asymptotic distribution of the statistic. However, since the focus of this article lies on detecting systematic forecasting bias rather than dealing with specific forms of estimation error, we consider the latter only under the expanding scheme with a "large" in-sample estimation window. Specifically, for forecasting models, we assume that we also observe R additional observations of y_t prior to $R+1$ that may be used as estimation window (note that the R in-sample observations also comprise H observations that are used to produce the initial out-of-sample forecast for period $R+1$), and that $T = R + P$ with $P/R \to 0$ as $P, R \to \infty$. This allows us to abstract from estimation error in the analysis and to focus on systematic features of the forecasts.[3]

The parametric models we consider in this article take the form $m(\mathbf{W}_{t-h}; \boldsymbol{\theta}^{\dagger}_{\tau_k,h})$, where $\mathbf{W}_{t-h}$ denotes a vector of predictor variable(s) and we assume for simplicity that $\mathbf{W}_{t-h}$ is a subset of $\mathbf{V}_{t-h}$. Moreover, $\boldsymbol{\theta}^{\dagger}_{\tau_k,h}$ is a population parameter vector that needs to be estimated in a first step, while the function $m(\mathbf{W}_{t-h}; \cdot)$ on the other hand is assumed to be a "smooth" function of the parameter vector in the sense of Assumption A6. For instance, $m(\mathbf{W}_{t-h}; \boldsymbol{\theta}^{\dagger}_{\tau_k,h})$ could itself take the form of a linear quantile regression model:

$$q_{t,h}\left(\tau_k|\mathbf{W}_{t-h}\right) = m(\mathbf{W}_{t-h}; \boldsymbol{\theta}^{\dagger}_{\tau_k,h}) = \mathbf{W}'_{t-h}\boldsymbol{\theta}^{\dagger}_{\tau_k,h}.$$

Alternatively, the model could also take the form of a nonlinear location scale model:

$$\begin{aligned} q_{t,h}\left(\tau_k|\mathbf{W}_{t-h}\right) &= m(\mathbf{W}_{t-h}; \boldsymbol{\theta}^{\dagger}_{\tau_k,h}) \\ &= m_{\mu}\left(\mathbf{W}_{t-h}; \boldsymbol{\theta}^{\dagger}_{h,\mu}\right) + \sigma\left(\mathbf{W}_{t-h}; \boldsymbol{\theta}^{\dagger}_{h,\sigma}\right) q_{t,h,\epsilon}\left(\tau_k\right), \end{aligned}$$

where $\boldsymbol{\theta}^{\dagger}_{\tau_k,h} = (\boldsymbol{\theta}^{\dagger\prime}_{h,\mu}, \boldsymbol{\theta}^{\dagger\prime}_{h,\sigma}, q_{t,h,\epsilon}\left(\tau_k\right))'$ and $q_{t,h,\epsilon}\left(\tau_k\right)$ denotes the unconditional τ_k quantile of the error term of the location scale model.

To accommodate both "forecasting methods" and "forecasting models," we adopt a more generic notation in what follows and let $\mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}^{\dagger}_{\tau_k,h})$ stand either for the vector stemming from a forecasting method or for the population vector of Mincer-Zarnowitz regressors stemming from a corresponding forecasting model. On the contrary, when forecasts have been generated from a model that has been estimated through the recursive (i.e., expanding) scheme using the first $t-h$ observations of the

[3] In addition, it also allows us to resample forecasts directly in the bootstrap.

sample, with $t = R+1, R+2, \ldots$, we write $\mathbf{X}_{\tau_k,t,h}(\widehat{\boldsymbol{\theta}}_{\tau_k,t,h})$ to denote the dependence on the estimated parameter vector $\widehat{\boldsymbol{\theta}}_{\tau_k,t,h}$. Note that even though we focus on the recursive scheme in light of the applications, the rolling estimation scheme whereby a window of the last R observations is used (running from $t-h-R+1$ to $t-h$) or the fixed scheme where the parameter vector is estimated only once, that is $\widehat{\boldsymbol{\theta}}_{\tau_k,t,h} = \widehat{\boldsymbol{\theta}}_{\tau_k,R+1-h,h}$, are equally compatible with our set-up and the assumptions below could be adapted in a straightforward manner. Finally, "forecasting methods" can be accommodated by assuming $\widehat{\boldsymbol{\theta}}_{\tau_k,t,h} = \boldsymbol{\theta}^{\dagger}_{\tau_k,h}$ almost surely.

Thus, to implement the test for the null hypothesis in (4) versus the alternative hypothesis, we first estimate the coefficient vector as

$$\widehat{\boldsymbol{\beta}}_h(\tau_k) = \begin{pmatrix} \widehat{\alpha}_h(\tau_k) \\ \widehat{\beta}_h(\tau_k) \end{pmatrix} = \arg\min_{\boldsymbol{b} \in \mathcal{B}} \frac{1}{P} \sum_{t=R+1}^{T} \rho_\tau \left(y_t - \mathbf{X}_{\tau_k,t,h}(\widehat{\boldsymbol{\theta}}_{\tau_k,t,h})' \boldsymbol{b} \right) \quad (5)$$

for each h and τ_k. With these estimates at hand, different possibilities to construct a suitable test statistic exist. More specifically, since the number of elements in $\mathcal{H}$ and $\mathcal{T}$ is finite, one option is to construct a Wald-type test based on the estimates in (5) together with a suitable estimator of the variance-covariance matrix. However, when interest lies in testing (4) against its complement for a larger number of quantile levels and horizons, constructing a Wald test involves estimating a large variance-covariance matrix, which can be difficult in practice and may lead to a poor finite sample performance. On the other hand, as we argue below, using a moment equality based test in combination with the nonparametric bootstrap does not suffer from this drawback. In fact, the moment equality framework can be extended easily to other set-ups which give rise to an even larger number of equalities (see Section 4).

To see the possibility of a moment equality based test, note that under H_0^{MZ} and the Assumptions A1 to A7 outlined below it holds that

$$\sqrt{P}\left(\widehat{\boldsymbol{\beta}}_h^{\dagger}(\tau_k) - \begin{pmatrix} \alpha_h^{\dagger}(\tau_k) \\ \beta_h^{\dagger}(\tau_k) \end{pmatrix} \right) \xrightarrow{d} N\Big(\mathbf{0}, \tau_k(1-\tau_k)\mathbf{J}_h(\tau_k)^{-1} \mathrm{E}\left[\mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}^{\dagger}_{\tau_k,h}) \mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}^{\dagger}_{\tau_k,h})' \right] \mathbf{J}_h(\tau_k)^{-1} \Big)$$

pointwise in h and τ_k, where the matrix $\mathbf{J}_h(\tau_k)$ is given by

$$\mathbf{J}_h(\tau_k) \equiv \mathrm{E}\left[f_{t,h}\left(\mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}^{\dagger}_{\tau_k,h})' \boldsymbol{\beta}^{\dagger}(\tau_k) \right) \mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}^{\dagger}_{\tau_k,h}) \mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}^{\dagger}_{\tau_k,h})' \right] \quad (6)$$

and $f_{t,h}(\cdot)$ is defined in Assumption A4. In fact, under H_0^{MZ}, in Section S3 of the supplement we establish the linear Bahadur representation:

$$\sqrt{P}\left(\widehat{\boldsymbol{\beta}}_h(\tau_k) - \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right) = \mathbf{J}_h(\tau_k)^{-1} \left(\frac{1}{\sqrt{P}} \sum_{t=R+1}^{T} \mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}^{\dagger}_{\tau_k,h}) \times \left(1\left\{ y_t \le \mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}^{\dagger}_{\tau_k,h})' \boldsymbol{\beta}^{\dagger}_{\tau_k,h}(\tau_k) \right\} - \tau_k \right) \right) + o_{\Pr}(1).$$

The above representation motivates the use of a moment equality type statistic for a test of autocalibration. Thus, define the set $\mathcal{C}^{\mathrm{MZ}} = \left\{ (h, \tau_k, j) : h \in \mathcal{H}, \tau_k \in \mathcal{T}, j \in \{1, 2\} \right\}$, and let $|\mathcal{C}^{\mathrm{MZ}}| = \kappa$ denote the cardinality of $\mathcal{C}^{\mathrm{MZ}}$, while $s = 1, \ldots, \kappa$ is a generic element from $\mathcal{C}^{\mathrm{MZ}}$. For the test statistic, define $\widehat{m}_s$ either as $\widehat{\alpha}_h(\tau_k)$ or as $(\widehat{\beta}_h(\tau_k) - 1)$ for a specific τ_k and h. A test statistic for the null hypothesis in (4) is then given by

$$\widehat{U}_{\mathrm{MZ}} = \sum_{s=1}^{\kappa} \left(\sqrt{P} \widehat{m}_s \right)^2 . \quad (7)$$

Note that the non-studentized statistic in (7) above does not require estimation of the asymptotic variance, and consequently will be non-pivotal as its asymptotic distribution does depend on the full variance-covariance matrix. Heuristically, under H_0^{MZ}, and the conditions imposed in Theorem 1:

$$\sqrt{P} \left(\begin{pmatrix} \widehat{\alpha}_1(\tau_1) \\ \widehat{\beta}_1(\tau_1) \\ \vdots \\ \widehat{\alpha}_H(\tau_K) \\ \widehat{\beta}_H(\tau_K) \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \right) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}), \quad (8)$$

where $\boldsymbol{\Sigma}$ is the asymptotic variance-covariance matrix of the empirical moment equalities scaled by $\sqrt{P}$, which is unknown in practice and depends on features of the data generating process (DGP). Of course, this nuisance parameter problem can be taken into account by using a suitable bootstrap procedure (e.g., Hansen, Lunde, and Nason 2011). In particular, we will generate bootstrap critical values using the moving block bootstrap (MBB) of Künsch (1989), whose formal validity for quantile regression with time series observations has been established by Gregory, Lahiri, and Nordman (2018). In doing so, we will resample the forecasts directly as the limiting distribution will be derived under the condition that $P/R \to \pi = 0$, implying that forecast estimation error does not feature into the asymptotic distribution of the test statistic (see West 1996). We do so to abstract from the dependence on a particular estimator, which would require further details about the underlying forecasting models.

We generate bootstrap samples of length P consisting of K_b blocks of length l such that $P = K_b l$. We draw the starting index I_j of each block $1, \ldots, K_b$, $\{I_j, I_{j+1}, \ldots, I_{j+l}\}$, from a discrete random uniform distribution on $[R+1, T-l]$. These indices are used to resample from $\left\{ y_t, \widehat{y}_{\tau,t,h} \right\}_{t=R+1}^{T}$ jointly for each $\tau = \tau_1, \ldots, \tau_K$ and $h = 1, \ldots, H$. This way we generate B bootstrap samples, each with $\left\{ y_t^b, \widehat{y}^b_{\tau,t,h} \right\}_{t=R+1}^{T}$ for all $\tau = \tau_1, \ldots, \tau_K$ and $h = 1, \ldots, H$. For each bootstrap sample, we construct bootstrap equivalents of (5) and then the corresponding bootstrap statistic:

$$\widehat{U}^b_{\mathrm{MZ}} = \sum_{s=1}^{\kappa} \left(\sqrt{P} (\widehat{m}^b_s - \widehat{m}_s) \right)^2 . \quad (9)$$

The critical value is then given by the $(1-\alpha)$ quantile of the empirical bootstrap distribution of $\widehat{U}^b_{\mathrm{MZ}}$ over B draws, say $c_{B,P,(1-\alpha)}$.

3.3. Assumptions and Asymptotic Validity

For the asymptotic validity of this procedure, we make the following assumptions:

A1: The outcome variable y_t is strictly stationary and β-mixing with the mixing coefficient satisfying $\beta(k) = O\left(k^{-\frac{\epsilon}{\epsilon-1}}\right) < \infty$ for $\epsilon > 1$.

A2: For all $h \in \mathcal{H}$, $\tau_k \in \mathcal{T}$, and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta})$ is strictly stationary, and satisfies the mixing condition from A1 as well as $\mathrm{E}\left[\left\|\mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta})\right\|^{2\epsilon+2}\right] < \infty$, where $\|\cdot\|$ denotes the Euclidean norm and $\boldsymbol{\Theta}$ is defined in A6. The distribution of $\mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta})$ is absolutely continuous with Lebesgue density.

A3: For every $\tau_k \in \mathcal{T}$ and $h \in \mathcal{H}$, assume that the parameter space of $\boldsymbol{\beta}_h(\tau_k)$, $\mathcal{B}$, is a compact and convex set. For every $\tau_k \in \mathcal{T}$ and $h \in \mathcal{H}$, the coefficient vector $\boldsymbol{\beta}^\dagger_h(\tau_k)$ from (3) lies in the interior of $\mathcal{B}$.

A4: For all $h \in \mathcal{H}$ and $\tau_k \in \mathcal{T}$, the conditional distribution function of y_t (given $\mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}^\dagger_{\tau_k,h})$), $F(\cdot|\mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}^\dagger_{\tau_k,h})) \equiv F_{t,h}(\cdot)$, admits a continuous Lebesgue density, $f(\cdot|\mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta})) \equiv f_{t,h}(\cdot)$, which is bounded away from zero and infinity for all u in $\mathcal{U} = \{u : 0 < F_{t,h}(u) < 1\}$ almost surely. For all h, the density $f_{t,h}(\cdot)$ is integrable uniformly over $\mathcal{U}$.

A5: For every $h \in \mathcal{H}$ and $\tau_k \in \mathcal{T}$, assume that the matrix $\boldsymbol{J}_h(\tau_k)$ defined in (6) is positive definite and that:

$$\mathrm{E}\left[\mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}^\dagger_{\tau_k,h})\left(1\left\{y_t \le \mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}^\dagger_{\tau_k,h})'\boldsymbol{\beta}^\dagger_{\tau_k,h}(\tau_k)\right\} - \tau_k\right)\right] = \mathbf{0}.$$

A6: Assume that $\boldsymbol{\Theta}$ is compact and that, for each $\tau_k \in \mathcal{T}$ and $h \in \mathcal{H}$, $\boldsymbol{\theta}^\dagger_{\tau_k,h}$ lies in its interior. For all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$ and t, it holds that:

$$\|\mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}_1) - \mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}_2)\| \le B(\mathbf{X}_{\tau_k,t,h})\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|$$

for some nonnegative, real-valued function $B(\mathbf{X}_{\tau_k,t,h})$ satisfying $\mathrm{E}[B(\mathbf{X}_{\tau_k,t,h})^2] < \infty$. In addition, assume that for every $h \in \mathcal{H}$ and $\tau_k \in \mathcal{T}$, the estimator $\widehat{\boldsymbol{\theta}}_{\tau_k,t,h}$ satisfies:

$$\sup_{t \ge R+1} \|\widehat{\boldsymbol{\theta}}_{\tau_k,t,h} - \boldsymbol{\theta}^\dagger_{\tau_k,h}\| = O_{\mathrm{Pr}}\left(\frac{1}{\sqrt{R}}\right).$$

A7: Assume that $R, P, l \to \infty$ as $T \to \infty$ with $P/R \to \pi = 0$ and $l/P \to 0$.

Assumption A1 imposes some mild restrictions on the time dependence of the data that are in turn linked to the existence and finiteness of corresponding moments in A2. On the other hand, the continuity of $\mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta})$ for any given τ_k, h, and $\boldsymbol{\theta}$ only serves the purpose to simplify some of the arguments in the proofs of Section S3 in the supplement, and could be relaxed at the expense of more cumbersome notation. Assumptions A3–A5 are required to derive the limiting distribution of the linear quantile regression estimator (e.g., Koenker and Xiao 2006; Galvao, Montes-Rojas, and Olmo 2011). In fact, A3 and A4 represent standard assumptions on the parameter space and the smoothness of the (conditional) distribution of y_t. A5 ensures asymptotic normality of the quantile regression estimator, with the moment condition representing the quantile equivalent of the well known orthogonality condition from ordinary least squares (Kim and White 2003). Assumption A6 on the other hand is only needed when the focus lies on forecasting models. Specifically, it places restrictions on the underlying parametric models, but is in fact compatible with commonly used location scale or linear quantile regression models that satisfy the Lipschitz condition in A6 and that can be estimated at rate $\sqrt{R}$. In particular, as Koenker and Xiao (2009) propose a two-step estimation procedure for linear GARCH models based on linear quantile regression, our set-up also comprises the latter type of models that are frequently used in finance applications. Finally, Assumption A7 governs the rates at which P and R as well as the block length l may grow to infinity. In particular, and in analogy to West (1996), we require $\pi = 0$ for estimation error to be ignorable asymptotically and to be able to resample directly from the forecasts (rather than to resample from the realized predictors). In turn, this allows us to focus on miscalibration as a structural feature of the models.

We are now ready to derive the asymptotic properties of the statistic under the null hypothesis:

Theorem 1. Assume that A1–A7 hold, and that $\boldsymbol{\Sigma} \in \mathbb{R}^{\kappa \times \kappa}$ is positive definite. Then under H_0^{MZ}:

$$\lim_{T,B \to \infty} \Pr\left(\widehat{U}_{\mathrm{MZ}} > c_{B,P,(1-\alpha)}\right) = \alpha.$$

Theorem 1 establishes the asymptotic size control of the moment equality test. It is easy to implement using the moving block bootstrap and performs very well in terms of finite sample size and power, which we assess through several simulations. These can be found in Section S1 of the supplementary material. There we provide two contrasting simulation set-ups in Sections S1.1 and S1.3 to match both the macroeconomic and financial applications below, which differ in sample sizes, target quantile levels and time series characteristics (we choose AR and GARCH processes as DGPs). In terms of block length choice, we find that values like $l = 10$ work well in financial applications with thousands of daily observations and a lower length like $l = 4$ is adequate in macroeconomic situations with only hundreds of monthly observations. However, the results do not change significantly when tweaking the block length. We also provide additional simulations in Section S1.2 of the supplement for augmented quantile Mincer-Zarnowitz test that we discuss next.

4. Extensions

In the following two subsections, we will describe two extensions of the test from Section 3, first to accommodate additional predictors $\mathbf{Z}_{t-h}$ in the Mincer-Zarnowitz regression to test different forms of optimality, and second to test for autocalibration of quantile forecasts for multiple time series simultaneously.

4.1. Augmented Quantile Mincer-Zarnowitz Test

Recalling the characterization of optimality relative to any information set $\mathcal{I}_{t-h} \subset \mathcal{F}_{t-h}$ from (1), it becomes clear that the Mincer-Zarnowitz set-up from the previous section may also be used to test stronger forms of optimality with respect to larger information sets than $\sigma(\widehat{y}_{\tau_k,t,h})$. More precisely, while H_0^{MZ} versus H_1^{MZ} is a test of autocalibration, it does not check if all available valuable information from $\mathcal{F}_{t-h}$ was incorporated into the forecasting model or taken into account by the forecaster.

We therefore suggest the idea of augmented quantile Mincer-Zarnowitz regressions, where a vector of additional regressors $\mathbf{Z}_{t-h} \in \mathcal{F}_{t-h}$ is added to the regression model in (2) to test for optimality relative to $\sigma(\widehat{y}_{\tau_k,t,h}, \mathbf{Z}_{t-h})$, see also Elliott and Timmermann (2016, chap. 15) for a discussion of augmented Mincer-Zarnowitz regressions in the context of mean forecasts.

That is, in analogy to the previous section, we again specify a linear quantile regression model for every horizon $h \in \mathcal{H}$ and $\tau_k \in \mathcal{T}$ as follows:

$$y_t = \alpha_h^\dagger(\tau_k) + \widehat{y}_{\tau_k,t,h}\beta_h^\dagger(\tau_k) + \mathbf{Z}'_{t-h}\boldsymbol{\gamma}_h^\dagger(\tau_k) + \varepsilon_{t,h}(\tau_k), \quad (10)$$

where with slight abuse of notation we use the same symbols as in the previous section for the first two regression coefficients as well as the error term, and suppress the possible dependence of the forecasts on some parameter vector $\boldsymbol{\theta}^\dagger$. If the coefficients of $\mathbf{Z}_{t-h}$ in the population augmented Mincer-Zarnowitz regression are nonzero, that is $\boldsymbol{\gamma}_h^\dagger(\tau_k) \neq \mathbf{0}$, there is valuable information in $\mathbf{Z}_{t-h}$ that has not been incorporated into the forecasts yet. As a result, those variables or a subset thereof should be included into the model to improve forecast accuracy.

Formally, the null hypothesis we test is given by:

$$H_0^{\text{AMZ}} : \{\alpha_h^\dagger(\tau_k) = 0\} \cap \{\beta_h^\dagger(\tau_k) = 1\} \cap \{\boldsymbol{\gamma}_h^\dagger(\tau_k) = \mathbf{0}\} \quad (11)$$

for all $h \in \mathcal{H}$ and $\tau_k \in \mathcal{T}$ versus $H_1^{\text{AMZ}} : \{\alpha_h^\dagger(\tau_k) \neq 0\}$ and/or $\{\beta_h^\dagger(\tau_k) \neq 1\}$ and/or $\{\boldsymbol{\gamma}_h^\dagger(\tau_k) \neq \mathbf{0}\}$ for at least some $h \in \mathcal{H}$ and $\tau_k \in \mathcal{T}$.

In contrast to a standard Mincer-Zarnowitz quantile regression, the augmented version requires a choice of variables from $\mathcal{F}_{t-h}$. These variables included in $\mathbf{Z}_{t-h}$ have to be chosen a priori and may, in some situations, suggest themselves naturally as illustrated in our macroeconomic application. In other situations, however, it might be hard to pick those variables from a potentially very large information set. In those cases, regularized or factor-augmented Mincer-Zarnowitz regressions could be used instead of (10). We leave this extension to future research and state the asymptotic size result of a moment equality based test of H_0^{AMZ} versus H_1^{AMZ}. Specifically, let $\widehat{\widetilde{m}}_s$ denote the quantile estimators $\widehat{\alpha}_h(\tau_k)$, $(\widehat{\beta}_h(\tau_k) - 1)$, or $\widehat{\boldsymbol{\gamma}}_h(\tau_k)$ for a specific quantile level τ_k and horizon h. Similarly, in analogy to before, let $\widetilde{\kappa}$ denote the total (finite) number of moment equalities to be tested. The corresponding test statistic, $\widehat{U}_{\text{AMZ}}$, is given by

$$\widehat{U}_{\text{AMZ}} = \sum_{s=1}^{\kappa} \left(\sqrt{P}\widehat{\widetilde{m}}_s\right)^2. \quad (12)$$

Finally, define the "augmented" covariate and coefficient vectors:

$$\widetilde{\mathbf{X}}_{\tau_k,t,h}(\boldsymbol{\theta}^\dagger_{\tau_k,h}) = (\mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}^\dagger_{\tau_k,h})', \mathbf{Z}'_{t-h})'$$

and $\widetilde{\boldsymbol{\beta}}_h^\dagger(\tau_k) = (\alpha_h^\dagger(\tau_k), \beta_h^\dagger(\tau_k), \boldsymbol{\gamma}_h^\dagger(\tau_k)')'$, respectively, the latter with parameter space $\widetilde{\mathcal{B}}$. We obtain the following result for the test defined by (12):

Corollary 1. Assume that A1–A7 hold with $\mathbf{X}_{\tau_k,t,h}(\cdot)$, $\boldsymbol{\beta}_h^\dagger(\tau_k)$, and $\mathcal{B}$ replaced by $\widetilde{\mathbf{X}}_{\tau_k,t,h}(\cdot)$, $\widetilde{\boldsymbol{\beta}}_h^\dagger(\tau_k)$, and $\widetilde{\mathcal{B}}$, respectively. Moreover, assume that

$$\operatorname*{plim}_{P\to\infty} \text{var}\left[\sqrt{P}\begin{pmatrix}\widehat{\widetilde{m}}_1 \\ \vdots \\ \widehat{\widetilde{m}}_{\overline{\kappa}}\end{pmatrix}\right] \equiv \widetilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{\widetilde{\kappa}\times\widetilde{\kappa}}$$

is positive definite, where var $[\cdot]$ denotes the variance operator. Then under H_0^{AMZ}:

$$\lim_{T,B\to\infty} \Pr\left(\widehat{U}_{\text{AMZ}} > c_{B,P,(1-\alpha)}\right) = \alpha.$$

4.2. Multivariate Quantile Mincer-Zarnowitz Test

While Section 3 establishes a test for autocalibration of quantile forecasts of a single time series y_t, researchers may sometimes be interested in testing this property across several time series $\mathbf{y}_t = (y_{1,t}, \ldots, y_{G,t})'$ using an array of h step ahead τ_k-level forecasts $\left(\widehat{\mathbf{y}}_{\tau,t,h}\right)_{\tau=\tau_1,\ldots,\tau_K, h=1,\ldots,H}$, where $\widehat{\mathbf{y}}_{\tau_k,t,h} = (\widehat{y}_{1,\tau_k,t,h}, \ldots, \widehat{y}_{G,\tau_k,t,h})'$, where $h \in \mathcal{H}$, $\tau_k \in \mathcal{T}$, and finite $G \in \mathbb{N}$. For instance, we may be interested in testing for forecast autocalibration jointly across different industries or sectors; components of GDP growth like consumption or export growth; or different macro series like in our application in Section 5.2.

We now sketch the extension of the autocalibration test to such a multivariate set-up. To this end, consider again the following linear quantile regression model:

$$y_{i,t} = \alpha_{i,h}^\dagger(\tau_k) + \widehat{y}_{i,\tau_k,t,h}\beta_{i,h}^\dagger(\tau_k) + \varepsilon_{i,t,h}(\tau_k),$$
$$h \in \mathcal{H}, \quad \tau_k \in \mathcal{T}, \quad i = 1, \ldots, G. \quad (13)$$

Here, for each group $i \in \{1, \ldots, G\}$, the coefficient vector $\boldsymbol{\beta}_{i,h}^\dagger(\tau_k) = (\alpha_{i,h}^\dagger(\tau_k), \beta_{i,h}^\dagger(\tau_k))'$ is defined as

$$\boldsymbol{\beta}_{i,h}^\dagger(\tau_k) = \arg\min_{\boldsymbol{b}_i\in\mathcal{B}} \text{E}\left[\left(\rho_{\tau_k}\left(y_{i,t} - \mathbf{X}_{i,\tau_k,t,h}(\boldsymbol{\theta}^\dagger_{i,\tau_k,h})'\boldsymbol{b}_i\right)\right)\right] \quad (14)$$

with $\mathbf{X}_{i,\tau_k,t,h}(\boldsymbol{\theta}^\dagger_{i,\tau_k,h})' = (1, \widehat{y}_{i,\tau_k,t,h})'$.[4] Therefore, the set-up in (13) allows for miscalibration also at the individual time series level (e.g., industry or sector) if for some i, h, and τ_k it holds that $\alpha_{i,h}(\tau_k) \neq 0$ and/or $\beta_{i,h}(\tau_k) \neq 1$, respectively.

The sample analogue of (14), using the evaluation sample of observations $\{y_t\}_{t=R+1}^T$ and array-valued forecasts $\widehat{y}_{i,\tau,t,h}$ for each i is given by

$$\widehat{\boldsymbol{\beta}}_{i,h}(\tau_k) = \arg\min_{\boldsymbol{b}_i\in\mathcal{B}} \frac{1}{P}\sum_{s=R+1}^{T}\left(\rho_\tau\left(y_{i,s} - \mathbf{X}_{i,\tau_k,s,h}(\widehat{\boldsymbol{\theta}}_{i,\tau_k,s,h})'\boldsymbol{b}_i\right)\right). \quad (15)$$

As in Section 3, we are interested in testing the composite null hypothesis:

$$H_0^{\text{MMZ}} : \{\alpha_{i,h}(\tau_k) = 0\} \cap \{\beta_{i,h}(\tau_k) = 1\} \quad (16)$$

for all $h \in \mathcal{H}$, $\tau_k \in \mathcal{T}$, $i = 1, \ldots, G$, versus $H_1^{\text{MMZ}} : \{\alpha_{i,h}(\tau_k) \neq 0\}$ and/or $\{\beta_{i,h}(\tau_k) \neq 1\}$ for at least some h, τ_k, and i. Note that, since G is finite, for a given horizon h and quantile level τ_k (with A1–A7 adapted to hold for $y_{i,t}$ and $\mathbf{X}_{i,\tau_k,t,h}(\boldsymbol{\theta}^\dagger_{i,\tau_k,h})$, $i = 1, \ldots, G$), the estimator in (15) is consistent for $\boldsymbol{\beta}_{i,h}^\dagger(\tau_k)$ for each i. The test statistic for the null hypothesis in (16) versus its complement is therefore given by $\widehat{U}_{\text{MMZ}} = \sum_{s=1}^{\overline{\kappa}}\left(\sqrt{P}\widehat{\overline{m}}_s\right)^2$, where either $\widehat{\overline{m}}_s = \widehat{\alpha}_{i,h}(\tau_k)$ or $\widehat{\overline{m}}_s = \widehat{\beta}_{i,h}(\tau_k) - 1$, respectively, and in a similar way as above, $\overline{\kappa}$ denotes the total number of moment conditions which in this case accounts for the G different series being used in the test.

[4] Note that we use $\boldsymbol{\theta}_i^\dagger$ to denote the possible dependence of the forecast for series i on a forecasting model.

In order to construct a suitable bootstrap statistic in analogue to Section 3.2, we construct bootstrap analogues $\widehat{\boldsymbol{\beta}}^{b}_{i,h}(\tau_k)$ of (15) from bootstrap samples of length $P = K_b l$ from K_b blocks of length l by resampling again from the series of forecast-observation pairs, where the forecasts in this case are array-valued. The bootstrap procedure does not just take the horizon, but also the group structure as given, which ensures that the dependence of the original data across horizons h as well as across series $i = 1, \ldots, G$ is maintained. More precisely, we again draw the starting index I_j of each block of forecasts and observations $1, \ldots, K_b$ from a discrete random uniform distribution on $[R + 1, T - l]$. These indices are used to resample from $\left\{y_t, \left(\widehat{\mathbf{y}}_{\tau,t,h}\right)_{\tau=\tau_1,\ldots,\tau_K,h=1,\ldots,H}\right\}_{t=R+1}^{T}$. This way we generate B bootstrap samples, each with $\left\{y_t^b, \left(\widehat{\mathbf{y}}^b_{\tau,t,h}\right)_{\tau=\tau_1,\ldots,\tau_K,h=1,\ldots,H}\right\}_{t=R+1}^{T}$.

For each $i = 1, \ldots, G$, we then construct a corresponding bootstrap estimator given by

$$\widehat{\boldsymbol{\beta}}^{b}_{i,h}(\tau_k) = \arg\min_{\boldsymbol{b}_i \in \mathcal{B}} \frac{1}{P} \sum_{s=R+1}^{T} \left(\rho_\tau \left(y^b_{i,s} - \mathbf{X}_{i,\tau_k,s,h}(\widehat{\boldsymbol{\theta}}^{b}_{i,\tau_k,s,h})' \boldsymbol{b}_i\right)\right).$$

The final bootstrap statistic becomes $\widehat{U}^b_{\text{MMZ}} = \sum_{s=1}^{\overline{\kappa}} \left(\sqrt{T}(\widehat{\overline{m}}^b_s - \widehat{\overline{m}}_s)\right)^2$, where $\widehat{\overline{m}}^b_s$ is equal to $\widehat{\alpha}^b_{i,h}(\tau_k)$ or $\widehat{\beta}^b_{i,h}(\tau_k) - 1$, respectively. Constructing critical values on the basis of $\widehat{U}^b_{\text{MMZ}}$, $b = 1, \ldots, B$, as in Section 3.2, the following corollary holds:

Corollary 2. Assume that Assumptions A1–A7 hold with y_t, $\mathbf{X}_{\tau_k,t,h}(\boldsymbol{\theta}^{\dagger}_{\tau_k,h})$, and $\widehat{\boldsymbol{\theta}}_{\tau_k,t,h}$ replaced by $y_{i,t}$, $\mathbf{X}_{i,\tau_k,t,h}(\boldsymbol{\theta}^{\dagger}_{i,\tau_k,h})$, and $\widehat{\boldsymbol{\theta}}_{i,\tau_k,t,h}$, respectively, for every $i = 1, \ldots, G$. Moreover, assume that:

$$\operatorname*{plim}_{P\to\infty} \operatorname{var}\left[\sqrt{P}\begin{pmatrix} \widehat{\overline{m}}_1 \\ \vdots \\ \widehat{\overline{m}}_{\overline{\kappa}} \end{pmatrix}\right] \equiv \overline{\boldsymbol{\Sigma}} \in \mathbb{R}^{\overline{\kappa}\times\overline{\kappa}},$$

is positive definite, where var $[\cdot]$ denotes the variance operator. Then, under H_0^{MMZ}:

$$\lim_{T,B\to\infty} \Pr\left(\widehat{U}_{\text{MMZ}} > c_{B,P,(1-\alpha)}\right) = \alpha.$$

5. Empirical Applications

In this section, we provide two empirical applications. The first is a finance example applying the MZ test to test the optimality of GARCH predictions of the tail quantiles of financial returns. The second application uses the MZ test extensions to assess the optimality of GaR forecasts made across a range of U.S. macroeconomic variables.

5.1. Empirical Application 1: Financial Returns

Forecasts of lower tail quantiles of returns distributions (or upper tail quantiles of loss distributions) play an important role in financial risk management as the most prominent risk measures are either themselves tail quantiles or defined in terms of tail quantiles (see He, Kou, and Peng (2022) for a recent overview). In particular, the VaR at level τ is just the τ-quantile of the returns y_t, $VaR(\tau) = q(\tau)$, where usually τ is chosen to be either 0.05 or 0.01. Expected shortfall as well as median shortfall are also defined in terms of quantiles of the return distribution and our multi-quantile evaluation framework can be useful in the evaluation of those risk measures as well (see Section S5.1 of the supplement for a discussion). Producing and backtesting VaR forecasts is thus a central task in financial risk management. Therefore, as discussed in the introduction, the majority of contributions to quantile forecast optimality testing was motivated by this problem. However, the literature focused on single-quantile forecasts, while it may be of interest to check VaR at both the 0.01 and the 0.05 level, or even a grid of several VaR levels to approximate the whole tail distribution. In addition, note that risk management requires forecasts of risk measures over multiple horizons, for example for one day ahead and cumulative losses over the next 10 trading days. Nevertheless, extant evaluation methods focus on a single horizon (with the noteworthy exception of Barendse, Kole, and van Dijk (2023) who also consider multi-horizon evaluation of VaR and expected shortfall) and consequently one-day-ahead forecasts are usually evaluated. Our tests solve those problems as they enable joint evaluation of VaR forecasts over multiple levels and horizons.

To illustrate the use of the Mincer-Zarnowitz test for the evaluation of forecasts of financial risk measures, we apply it to multi-horizon, multi-quantile forecasts for daily S&P 500 returns. We consider horizons from $h = 1$ through $h = 10$ in terms of trading days and three quantile levels $\tau \in \{0.01, 0.025, 0.05\}$. The classic model for return volatility and VaR forecasting is the GARCH(1,1) model (Bollerslev 1986). As no closed-form formula for multi-period-ahead GARCH quantile forecasts is available, except for the case of Gaussian innovations, we use the GARCH bootstrap of Pascual, Romo, and Ruiz (2006). It draws standardized residuals from the estimated one-period-ahead model to simulate draws multiple periods in the future, from which quantiles can be obtained.[5] We choose student-t errors for the estimation of the model.

Our sample consists of daily S&P 500 returns from January 3rd 2000 to June 27th 2022, amounting to 5634 observations.[6] We use recursive pseudo-out-of-sample forecasting with an initial estimation window of size 3000 for $h = 10$ or, in other words, $R = 3009$, leading to an evaluation sample of size $P = 2625$. Figure S1 in Section S5.2 of the supplementary material displays the one-day ahead forecasts for the three quantiles and the realizations. The forecasts for the other horizons look very similar, but are expectedly a bit wider.

We first use our Mincer-Zarnowitz test over the three quantiles, $\mathcal{T} = \{0.01, 0.025, 0.05\}$, and ten horizons, $\mathcal{H} = \{1, \ldots, 10\}$. We use $B = 1000$ bootstrap draws and a block length of $l = 10$. Table 1 presents the results. With a p-value of 0.022 there is clear evidence against the null of autocalibration. Alternative block length choices of $l = 5$ and $l = 20$ lead to similar p-values (0.030

[5] We use the implementation of the GARCH bootstrap from the `rugarch` package in R (Ghalanos 2022).

[6] Data taken from the Oxford-Man Realized Library: *https://realized.oxford-man.ox.ac.uk/data/download* [Last accessed: 05/07/22]

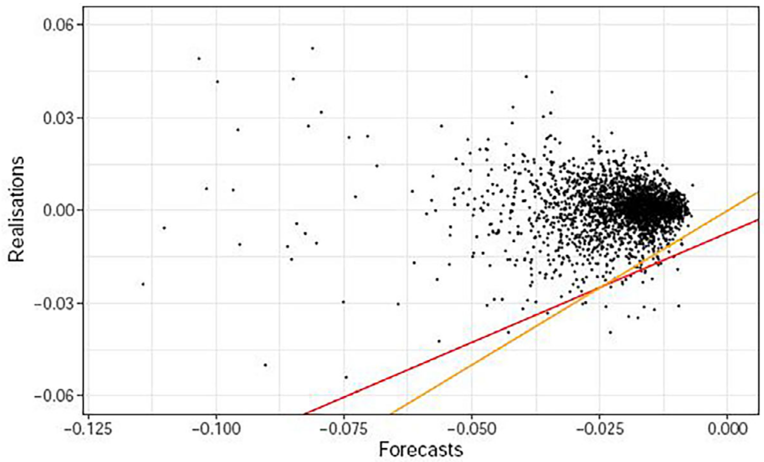

Figure 1. Scatterplot of Forecast-Realization Pairs with Mincer-Zarnowitz Regression Line (red) and Diagonal (orange) for $h = 1$ and $\tau = 0.01$

Table 1. Mincer-Zarnowitz test results, finance application.

Stat	90%	95%	99%	p-value
8932.46	5375.701	7072.245	11224.643	0.022

Table 2. Individual contributions to test statistic, Mincer-Zarnowitz test, finance application.

	$\tau = 0.01$	$\tau = 0.025$	$\tau = 0.05$	Sum
$h = 1$	221.693	51.298	17.953	290.944
$h = 2$	399.867	182.336	63.023	645.225
$h = 3$	263.330	193.565	95.534	552.429
$h = 4$	460.803	252.513	154.109	867.424
$h = 5$	559.772	248.379	105.181	913.332
$h = 6$	607.838	336.089	235.282	1179.210
$h = 7$	201.185	319.621	250.535	771.341
$h = 8$	344.977	383.845	438.533	1167.356
$h = 9$	350.399	342.559	405.926	1098.884
$h = 10$	449.902	586.011	410.401	1446.314
Sum	3859.767	2896.216	2176.476	8932.460

and 0.008) which is promising in that the results are insensitive to block length.

As the test statistic from (7) can directly be interpreted as an empirical distance from the null, consisting of scaled (by $\sqrt{P}$), squared deviations of all the Mincer-Zarnowitz regression coefficients from their values under the null, we can also look at the individual contributions to this statistic from single quantiles and single horizons or single quantile-horizon combinations, displayed in Table 2. From this table a clear picture emerges. The outer quantiles and the longer forecast horizons contribute more to the test statistic, and thus show stronger evidence for miscalibration. Since risk management is typically concerned about the performance of a certain risk model such as the GARCH(1,1) across a range of quantile levels or horizons, this also demonstrates that a common practice to evaluate those models only for a specific choice of the latter may lead to incorrect conclusions about the overall performance of the prediction model.[7]

The tests may also convey information about how models could be improved by a closer look at the Mincer-Zarnowitz regression lines themselves. For instance, using $h = 1$ and $\tau = 0.01$ as an illustrative example since all estimated intercepts are negative and all slopes of the regression lines less than one (see Section S5.2 of the supplementary material), Figure 1 shows the scatterplot of forecast-observation pairs alongside the estimated Mincer-Zarnowitz regression line and the diagonal. The latter represents the population regression line under H_0^{MZ}, in other words when $\alpha_1^\dagger(0.01) = 0$ and $\beta_1^\dagger(0.01) = 1$, respectively. The discrepancy between the Mincer-Zarnowitz regression line and the diagonal thus suggests that the forecasts are in fact miscalibrated. Contrasting the two, it becomes clear that in calmer times (when the forecasts and realizations are less extreme, that is closer to 0) the GARCH(1,1) forecasts tend to under-predict the actual risk, in other words the quantile forecasts are not extreme enough, while in more volatile times the forecasts tend to overestimate risk.

Finally, in Section S5.3 of the supplementary material, we examine the robustness of our results against various specification changes and examine other popular calibration tests designed for single horizon and quantile pairs. Specifically, we find that the above results do not change qualitatively when we alter the specification to a different estimation scheme (rolling) and to smaller estimation window sizes. Similarly, results remained also unchanged when omitting the COVID-19 period, or when experimenting with the GJR-GARCH model (Glosten, Jagannathan, and Runkle 1993). Finally, when looking at the autocalibration tests of Engle and Manganelli (2004) with the respective quantile forecast as regressor and the test of Christoffersen (1998), we find that the former provides p-values that are very close to the individual level p-values of the Mincer-Zarnowitz test, while this is not the case for the test of Christoffersen (1998) which tests different implications of full optimality.

5.2. Empirical Application 2: U.S. Macro Series

In this section we use our tests to explore the optimality of model-based forecasts of various U.S. macroeconomic series. The analysis of quantile forecasts for macroeconomic series has

[7] In fact, Table S8 in Section S5.2 of the supplement, which contains the p-values for individual autocalibration tests at given values τ and h, illustrates that such "telescoping'" practice may indeed be misleading as there is no strong evidence against autocalibration from some individual level quantile horizon combinations.

become widespread since studies like Manzan (2015). More recently, the GaR literature has emerged to provide a tool to monitor downside risk to economic growth using quantile predictions. This approach typically analyses quarterly real GDP growth using financial conditions indicators (see Adrian, Boyarchenko, and Giannone 2019), and has been subsequently applied to other quarterly macro series like employment and inflation by Adams et al. (2021).

However, in spite of the increasing interest in quantile forecasting in macroeconomics, none of these papers subject their models to the type of forecast optimality test we develop in this article. We aim to fill this gap in the empirical literature, applying our tests to shed light on the optimality of commonly-used models in predicting various macro series.

Instead of using quarterly data we propose the use of monthly variables (also used recently in similar contexts by Corradi, Fosten, and Gutknecht 2023) and we will focus on the same four target variables analyzed in Manzan (2015). These series, all transformed to stationarity using the growth rate, are the Consumer Price Index for All Urban Consumers (CPIAUCSL), Industrial Production: Total Index (INDPRO), All Employees, Total Nonfarm (PAYEMS) and Personal Consumption Expenditures Excluding Food and Energy (Chain-Type Price Index) (PCEPILFE).[8] These series are very close in nature to the quarterly series analyzed in Adams et al. (2021) and will be regressed on an autoregressive term and the Chicago Fed National Financial Conditions Index (NFCI) as in Adrian, Boyarchenko, and Giannone (2019).

Specifically, we use the direct forecasting scheme to generate quantile forecasts at quantile levels τ_k for $k = 1, \ldots, K$ and horizons $h = 1, \ldots, H$ as follows:

$$\widehat{y}_{\tau_k,t,h} = \widehat{\gamma}_{0,h,t}(\tau_k) + \widehat{\gamma}_{1,h,t}(\tau_k) y_{t-h} + \widehat{\gamma}_{2,h,t}(\tau_k) x_{t-h} \qquad (17)$$

where y_{t-h} is the autoregressive term corresponding to one of the four target variables mentioned above and x_{t-h} is the NFCI. The parameter estimates are obtained by the standard quantile regression estimator and are indexed both by τ_k and h to denote that a separate quantile regression is run at each quantile and horizon as in the direct scheme, as well as by t as the forecasts are generate in a pseudo out-of-sample fashion as mentioned below. In essence, (17) boils down to a forecast made by a quantile autoregressive distributed lag (QADL) model (Galvao Jr., Montes-Rojas, and Park 2013) using the direct forecasting scheme.

The data series span the period 1984M1 to 2019M12, giving a total number of $T = 432$ monthly observations. We use the recursive out-of-sample scheme and split the sample into equal portions for the initial estimation sample and the evaluation sample, $R = P = 216$. This gives an evaluation sample size, P, around the middle of the range of Monte Carlo simulations in Section S1 of the supplement. In making forecasts using (17) we will use horizons $h = 1, \ldots, 12$ and quantile levels $\tau_k \in \{0.1, 0.25, 0.5\}$. The use of these quantile levels allows us to focus on the left part of the distribution, as is common in GaR studies such as Adams et al. (2021), but also includes the median as an important case of predicting the center of the distribution. For the bootstrap implementation we use $B = 1000$ bootstrap draws and employ a block length of $l = 4$ as this is seen to work well in the simulation study in the supplement.

Table 3. Mincer-Zarnowitz test results.

	Stat	90%	95%	99%	p-value
Joint	38264.280	28908.454	45259.085	86531.304	0.067
CPIAUCSL	18269.966	18033.852	32452.813	66594.353	0.099
INDPRO	4258.078	7578.204	11224.918	24413.160	0.222
PAYEMS	871.704	1574.085	2060.305	4994.712	0.308
PCEPILFE	14864.532	2316.907	2792.387	3678.394	0.000

Table 4. Augmented Mincer-Zarnowitz test results.

	Stat	90%	95%	99%	p-value
CPIAUCSL	21984.030	19794.203	29896.138	57657.304	0.085
INDPRO	5194.690	8722.551	12596.841	27604.813	0.224
PAYEMS	723.354	1494.399	2011.985	4470.360	0.350
PCEPILFE	15648.207	2455.174	2938.071	3801.048	0.000

The results in Table 3 display the results of the Mincer-Zarnowitz tests for autocalibration, with further graphical insight into the behavior of the out-of-sample predictions given in Sections S6.1 and S6.2 of the supplement. We first analyze the joint Mincer-Zarnowitz test ("Joint") which works on multiple time series, as described above, where in this context we have $G = 4$ target variables and we jointly test for autocalibration across all series to avoid the multiple testing problem. The results in the first row of Table 3 show that there is indeed some evidence against autocalibration when looking across all four macro series. The p-value of 0.07 indicates that there is evidence at the 10% significance level that the QADL-type model does not produce well-calibrated forecasts jointly across these four series, for forecast horizons $h = 1, \ldots, 12$ and quantile levels $\tau_k \in \{0.1, 0.25, 0.5\}$.

With this in mind, it is useful to dig further into the individual series to see which of them are likely to be causing the rejection of the joint null of autocalibration. The remainder of Table 3 displays the Mincer-Zarnowitz test when performed individually for each series. For the two real series, industrial production and employment, we see little evidence against the null. On the other hand, for the two price-type series we see somewhat different results with a clear rejection in the case of PCEPILFE and a p-value just under 10% for CPIAUCSL. This suggests that the QADL-type approach suggested by Adrian, Boyarchenko, and Giannone (2019) does indeed appear appropriate for real macroeconomic series but less-so for price series.[9]

One final exercise we perform is to apply the augmented MZ test where we use additional predictors in the MZ regression. Table 4 displays the results for each of the four series above, where in each case the remaining three variables were used as the augmenting regressors. This serves as a simple check to see if

[8] All series in the study are taken from the Federal Reserve Economic Data (FRED). Url: *https://fred.stlouisfed.org/* [Last accessed: 08/03/22]

[9] In Section S6.3 of the supplement, we also try to isolate the specific horizons and quantile levels that contribute the most to the rejection. Our findings suggest that for PCEPILFE and CPIAUCSL the smallest contribution to the statistic comes from quantile level $\tau_k = 0.5$, while the $\tau_k = 0.1$ quantile level contributes more substantially, even though no systematic conclusion can be drawn beyond $h = 1$. This seems to indicate that further study should consider investigating the types of series which might deliver better-calibrated predictions in the far-left tail of the distribution of inflation-type series.

any of these other variables would have been able to improve the forecasts if they were added to the forecasting model, especially for the real variables for which the weaker null of autocalibration was not rejected. However, the results in Table 4 are similar to the non-augmented version of the test. As expected, the stronger null is rejected as well for the inflation type series, which already showed rejections for the weaker null of autocalibration. More interestingly, for the real variables, we still get no rejections. This suggests that we are not able to improve these forecasts by the addition of inflation type variables to the forecasting model.

6. Conclusion

This article deals with the absolute evaluation of quantile forecasts in situations where predictions are made over multiple horizons and possibly multiple quantile levels. We propose multi-horizon, multi-quantile tests for optimality by employing quantile Mincer-Zarnowitz regressions and a moment equality framework with a bootstrap methodology which avoids the estimation of a large covariance matrix. The main quantile Mincer-Zarnowitz test is of the null hypothesis of autocalibration, which is a fundamental property of forecast consistency. We also provide two extensions. The first extension tests a stronger null hypothesis, which allows us to add further important variables to the information set with respect to which optimality is tested. This augmented quantile Mincer-Zarnowitz test thus makes it possible to examine if the information contained in those variables was used optimally by the forecaster. The second extension is a multivariate quantile Mincer-Zarnowitz test and allows us to check autocalibration of forecasts for multiple time series at possibly multiple horizons and quantiles.

Our tests allow for an overall decision about the quality of a forecasting approach, whether it is a single model used over multiple horizons and quantiles or a mix of different models and expert judgment employed by an institution. Crucially, it avoids the multiple testing problem inherent to most practical situations, where many forecasts are made over horizons, quantiles or multiple variables. Importantly, our testing framework is constructive in that it does not only provide a formal procedure to reach this overall decision, but may also provide valuable feedback about possible weaknesses of the forecasting approach under consideration and how it could be improved.

There are many possible future avenues arising from our work, for instance the evaluation of distributional or probabilistic forecasts (Gneiting and Katzfuss 2014). Since these distributional forecasts are considered quantile calibrated when the corresponding quantile forecasts for all quantiles are autocalibrated, one future extension of our work may look into optimality testing across a growing (with the sample size) number of quantiles.

Supplementary Materials

The supplementary materials contain additional Monte Carlo simulations, proofs for all theoretical results of the paper, an additional monotonicity test, as well as supplementary empirical results and figures for the applications.

Acknowledgments

We are grateful to Maximo Camacho, Mike Clements, Laura Coroneo, Marc Hallin, José Olmo, Emese Lazar, Adam Lee, and Xiaohan Xue for helpful comments on the paper as well as participants of the 12th ECB Conference on Forecasting Techniques (2023, Frankfurt), the Tenth Italian Congress of Econometrics and Empirical Economics 2023 (Cagliari), the SAEe 2022 in Valencia, the Statistical Week 2022 (Münster) and 2021 (virtual), the International Symposium on Forecasting 2021 (virtual), and seminar attendees at the University of Hagen, the University of Hohenheim and Goethe University Frankfurt, 2023, and at the ICMA Centre, University of Reading, 2022.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

Marc-Oliver Pohle is grateful for support by the Klaus Tschira Foundation, Germany.

ORCID

Marc-Oliver Pohle http://orcid.org/0000-0002-0020-3604

References

Adams, P. A., Adrian, T., Boyarchenko, N., and Giannone, D. (2021), "Forecasting Macroeconomic Risks," *International Journal of Forecasting*, 37, 1173–1191. [10]

Adrian, T., Boyarchenko, N., and Giannone, D. (2019), "Vulnerable Growth," *American Economic Review*, 109, 1263–1289. [1,2,10]

Andrews, D., and Soares, G. (2010), "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157. [1,2]

Antolin Diaz, J., Drechsel, T., and Petrella, I. (2021), "Advances in Nowcasting Economic Activity," *Mimeo*. [1]

Barendse, S., Kole, E., and van Dijk, D. (2023), "Backtesting Value-at-Risk and Expected Shortfall in the Presence of Estimation Error," *Journal of Financial Econometrics*, 21, 528–568. [8]

Bollerslev, T. (1986), "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics*, 31, 307–327. [8]

Brownlees, C. T., and Souza, A. (2021), "Backtesting Global Growth-at-Risk," *Journal of Monetary Economics*, 118, 312–330. [1]

Carriero, A., Clark, T. E., and Marcellino, M. (2020), "Nowcasting Tail Risks to Economic Activity with Many Indicators," *Federal Reserve Bank of Cleveland, working paper no. 20-13*. [1]

Christoffersen, P. F. (1998), "Evaluating Interval Forecasts," *International Economic Review*, 39, 841–862. [2,9]

Clements, M. P. (2022), "Forecaster Efficiency, Accuracy, and Disagreement: Evidence Using Individual-Level Survey Data," *Journal of Money, Credit and Banking*, 54, 537–568. [2]

Corradi, V., Fosten, J., and Gutknecht, D. (2023), "Conditional Quantile Coverage: An Application to Growth-at-Risk," *Journal of Econometrics*, 236, 105490. [2,10]

Elliott, G., and Timmermann, A. (2016), *Economic Forecasting*, Princeton: Princeton University Press. [7]

Engle, R. F., and Manganelli, S. (2004), "CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles," *Journal of Business & Economic Statistics*, 22, 367–381. [2,9]

Escanciano, J. C., and Olmo, J. (2010), "Backtesting Parametric Value-at-Risk With Estimation Risk," *Journal of Business & Economic Statistics*, 28, 36–51. [2]

Escanciano, J. C., and Olmo, J. (2011), "Robust Backtesting Tests for Value-at-risk Models," *Journal of Financial Econometrics*, 9, 132–161. [2]

Ferrara, L., Mogliani, M., and Sahuc, J.-G. (2021), "High-Frequency Monitoring of Growth at Risk," *International Journal of Forecasting*, 38, 582–595. [1]

Fosten, J., and Gutknecht, D. (2020), "Testing Nowcast Monotonicity with Estimated Factors," *Journal of Business & Economic Statistics*, 38, 107–123. [2]

Gaglianone, W. P., Lima, L. R., Linton, O., and Smith, D. R. (2011), "Evaluating Value-at-Risk Models via Quantile Regression," *Journal of Business & Economic Statistics*, 29, 150–160. [1,2,4]

Galvao, A., Montes-Rojas, G., and Olmo, J. (2011), "Threshold Quantile Autoregressive Models," *Journal of Time Series Analysis*, 32, 253–267. [6]

Galvao Jr., A. F., Montes-Rojas, G., and Park, S. Y. (2013), "Quantile Autoregressive Distributed Lag Model with an Application to House Price Returns," *Oxford Bulletin of Economics and Statistics*, 75, 307–321. [10]

Ghalanos, A. (2022), *rugarch: Univariate GARCH Models*, R package version 1.4-9. [8]

Giacomini, R., and Komunjer, I. (2005), "Evaluation and Combination of Conditional Quantile Forecasts," *Journal of Business & Economic Statistics*, 23, 416–431. [2]

Giacomini, R., and Rossi, B. (2010), "Forecast Comparisons in Unstable Environments," *Journal of Applied Econometrics*, 25, 595–620. [4]

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993), "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks," *Journal of Finance*, 48, 1779–1801. [9]

Gneiting, T. (2011), "Making and Evaluating Point Forecasts," *Journal of the American Statistical Association*, 106, 746–762. [3]

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007), "Probabilistic Forecasts, Calibration and Sharpness," *Journal of the Royal Statistical Society*, Series B, 69, 243–268. [3]

Gneiting, T., and Katzfuss, M. (2014), "Probabilistic Forecasting," *Annual Review of Statistics and Its Application*, 1, 125–151. [11]

Gneiting, T., and Ranjan, R. (2013), "Combining Predictive Distributions," *Electronic Journal of Statistics*, 7, 1747–1782. [1,3]

Granger, C. W. (1969), "Prediction with a Generalized Cost of Error Function," *Journal of the Operational Research Society*, 20, 199–207. [3]

Gregory, K., Lahiri, S., and Nordman, D. (2018), "A Smooth Block Bootstrap for Quantile Regression with Time Series," *Annals of Statistics*, 46, 1138–1166. [5]

Hansen, P. R., Lunde, A., and Nason, J. M. (2011), "The Model Confidence Set," *Econometrica*, 79, 453–497. [5]

He, X. D., Kou, S., and Peng, X. (2022), "Risk Measures: Robustness, Elicitability, and Backtesting," *Annual Review of Statistics and Its Application*, 9, 141–166. [8]

Kim, T.-H., and White, H. (2003), "Estimation, Inference, and Specification Testing for Possibly Misspecified Quantile Regression," in *Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later*, Volume 17 of Advances in Econometrics, eds. T. Fomby and R. Carter Hill, pp. 107–132, Bingley: Emerald Group Publishing Limited. [6]

Koenker, R., and Xiao, Z. (2006), "Quantile Autoregression," *Journal of the American Statistical Association*, 101, 980–990. [6]

Koenker, R., and Xiao, Z. (2009), "Conditional Quantile Estimation for Generalized Autoregressive Conditional Heteroscedasticity Models," *Journal of the American Statistical Association*, 104, 1696–1712. [6]

Künsch, H. (1989), "The Jackknife and the Bootstrap for General Stationary Observations," *Annals of Statistics*, 17, 1217–1241. [5]

Manzan, S. (2015), "Forecasting the Distribution of Economic Variables in a Data-Rich Environment," *Journal of Business & Economic Statistics*, 33, 144–164. [2,10]

Mincer, J. A., and Zarnowitz, V. (1969), "The Evaluation of Economic Forecasts," in *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, pp. 3–46. NBER. [4]

Nolde, N., and Ziegel, J. F. (2017), "Elicitability and Backtesting: Perspectives for Banking Regulation," *Annals of Applied Statistics*, 11, 1833–1874. [2]

Pascual, L., Romo, J., and Ruiz, E. (2006), "Bootstrap Prediction for Returns and Volatilities in GARCH Models," *Computational Statistics & Data Analysis*, 50, 2293–2312. [2,8]

Patton, A., and Timmermann, A. (2012), "Forecast Rationality Tests based on Multi-Horizon Bounds," *Journal of Business & Economic Statistics*, 30, 1–17. [2]

Plagborg-Møller, M., Reichlin, L., Ricco, G., and Hasenzagl, T. (2020), "When is Growth at Risk?" *BPEA Conference Draft, Spring*. [1]

Pohle, M.-O. (2020), "The Murphy Decomposition and the Calibration-Resolution Principle: A New Perspective on Forecast Evaluation," arXiv preprint arXiv:2005.01835. [3]

Prasad, A., Elekdag, S., Jeasakul, P., Lafarguette, R., Alter, A., Feng, A. X., and Wang, C. (2019), "Growth at Risk: Concept and Application in IMF Country Surveillance," IMF Working Paper 19/36, International Monetary Fund. [1]

Quaedvlieg, R. (2021), "Multi-Horizon Forecast Comparison," *Journal of Business & Economic Statistics*, 39, 40–53. [2]

Romano, J., and Shaikh, A. (2010), "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78, 169–211. [1]

Tsyplakov, A. (2013), "Evaluation of Probabilistic Forecasts: Proper Scoring Rules and Moments," Available at SSRN 2236605. [1,3]

West, K. D. (1996), "Asymptotic Inference about Predictive Ability," *Econometrica*, 64, 1067–1084. [5,6]