Evaluating Quantile-Bounded and Expectile-Bounded Interval Forecasts

James W. Taylor

Saïd Business School

University of Oxford

International Journal of Forecasting, 2021, 37(2), 800-811.

Address for Correspondence:

James W. Taylor
Saïd Business School
University of Oxford
Park End Street
Oxford OX1 1HP, UK
Tel: +44 (0)1865 288927
Email: james.taylor@sbs.ox.ac.uk

Evaluating Quantile-Bounded and Expectile-Bounded Interval Forecasts

Abstract

In many different contexts, decision making is improved by the availability of probabilistic predictions. The accuracy of probabilistic forecasting methods can be compared using scoring functions, and insight provided by calibration tests. These tests evaluate the consistency of predictions with the observations. Our main agenda in this paper is interval forecasts and their evaluation. Such forecasts are usually bounded by two quantile forecasts. However, a limitation of quantiles is that they convey no information regarding the size of potential exceedances. By contrast, the location of an expectile is dictated by the whole distribution. This prompts us to propose expectile-bounded intervals. We provide interpretation, a consistent scoring function and a calibration test. Before doing this, we reflect on the evaluation of forecasts of quantile-bounded intervals and expectiles, and suggest extensions of previously proposed calibration tests in order to guard against strategic forecasting. We illustrate ideas using day-ahead electricity price forecasting.

Keywords: Quantiles; Expectiles; Intervals; Scoring Functions; Calibration Tests; Electricity Prices.

1. Introduction

Probabilistic forecasts enhance decision making. In some applications, a prediction of the full probability distribution is required (Gneiting and Katzfuss, 2014), while in others, a prediction of a quantile or expectile may be needed as a measure of tail risk (Ziegel, 2016), or as the optimal point forecast when an asymmetric loss function is appropriate (Gneiting, 2011). Often, a probabilistic prediction is summarised by an interval forecast, in order to provide a simple way to convey forecast uncertainty (see, for example, Makridakis et al., 2018).

The evaluation of forecast accuracy enables the user to understand the quality of a method and to compare methods. Our main agenda in this paper is the evaluation of interval forecasts, which are usually defined as being bounded by two quantile forecasts. However, a quantile has the disadvantage of conveying no information regarding the potential magnitude of exceedances beyond the quantile. By contrast, the location of an expectile is dictated by the whole distribution (Koenker, 2005). For example, a change to the tail of the distribution beyond a quantile will not change that quantile, but it will affect all the expectiles. As another example, a change to one tail of a distribution does not change the quantiles of the other tail, but it does affect all the expectiles. A practical implication of this is that increased asymmetry in a distribution can be more easily detected by monitoring an extreme expectile than an extreme quantile. The advantages of expectiles motivates us to propose expectile-bounded intervals as a new way to convey forecast uncertainty. Given that an interval is often presented with a point forecast for the mean, an expectile-bounded interval has appeal, as the mean is itself the central expectile. For a quantile-bounded interval, it would be more consistent to provide a point forecast for the median than the mean. For some distributions, a quantile-bounded interval may not even contain the mean. In our analysis of expectile-bounded intervals, we consider their interpretability, as we acknowledge that expectiles are not as well understood, or indeed as intuitive, as quantiles.

Our consideration of this new form of interval forecast, and the evaluation of its

accuracy, is our main contribution. However, as a basis for this, we first review forecast evaluation for quantiles, quantile-bounded intervals and expectiles.

Forecasts from different methods can be compared using scoring functions (scores). For example, the quantile regression loss function is a scoring function for a quantile. A scoring function is said to be *consistent* if it is minimised by the true value of the functional[1]. An *elicitable* functional is one for which there exists a consistent score. For quantiles and expectiles, Gneiting (2011) provides the full set of consistent scoring functions. Both quantiles and expectiles have been used as measures of risk in finance. A risk measure is *coherent* if it has a set of attractive properties, such as subadditivity, which means that the measure for a portfolio cannot be greater than the sum of the measure for the constituent parts of the portfolio (Artzner et al., 1999). Quantiles are not coherent. In fact, expectiles are the only elicitable and coherent risk measures (Ziegel, 2016). This provides further motivation for considering expectile-bounded intervals to summarise forecast uncertainty.

As a complement to scoring functions, calibration tests provide insight that can potentially be used to improve accuracy. For example, for a forecast of the α quantile, a test of conditional calibrated examines whether the conditional probability of an observation falling below the forecast is equal to α. Nolde and Ziegel (2017) present a framework for calibration testing. Previously proposed calibration tests can be presented in this framework, such as the quantile calibration test of Engle and Manganelli (2004). This includes a term to guard against a *strategic* forecasting process that manipulates the forecast purely to pass the test. Examples of strategic forecasting in other contexts are discussed by Lichtendahl et al. (2013), Olszewski (2015) and Taylor (2020), who considers strategic predictive distributions. Although strategic forecasts are exposed as poor by a consistent score, there are several reasons why it is important that calibration tests cannot be gamed. Firstly, a seemingly calibrated strategic forecaster may

[1] A functional is a mapping from a class of probability distributions to the real line, for which the mean, quantiles and expectiles are examples (Gneiting and Katzfuss, 2014).

be viewed as outperforming a competitor that has a better score but fails the calibration test. Secondly, a forecaster that behaves strategically to some extent, or for some of the time, may be competitive in terms of both a calibration test and score. Thirdly, forecasts are sometimes evaluated using only calibration. We show how strategic prediction is a concern for the existing tests of forecasts of quantile-bounded intervals and expectiles. To address this, we use the framework of Nolde and Ziegel (2017) to present new calibration tests. Throughout this paper, the new tests that we propose guard against the only strategic forecasts that we can envisage. However, we acknowledge that we cannot be sure that other strategic forecasts may exist.

Section 2 reviews scoring functions and calibration testing for quantiles. Section 3 focuses on quantile-bounded intervals, and presents a new calibration test. Section 4 considers the evaluation of expectile forecasts, and extends an existing calibration test to guard against strategic forecasting. Section 5 proposes expectile-bounded intervals, provides interpretation for this form of interval, and introduces a scoring function and a calibration test. Section 6 uses electricity price data to illustrate the different forms of interval forecast and their evaluation. Section 7 provides a brief simulation study. Section 8 summarises and concludes the paper.

2. Quantile Forecasts

In this section, we review scoring functions and calibration tests for quantile forecasts. This material is not new, but we present it here as useful background for future sections.

2.1. Scoring Functions for Quantile Forecasts

The most widely used consistent quantile score is the quantile regression loss function (see Koenker and Machado, 1999; Taylor, 1999). We refer to it as the *quantile score*, and present it as follows:

$$S_\alpha^q\left(q_t(\alpha), y_t\right) = \left(\alpha - I\left\{y_t \le q_t(\alpha)\right\}\right)\left(y_t - q_t(\alpha)\right) \tag{1}$$

where y_t is the observation in period t, $q_t(\alpha)$ is the quantile for probability level α, and $I\{\cdot\}$ is the indicator function. Gneiting (2011) gives the general form of consistent scoring function for a quantile.

2.2. Calibration for Quantile Forecasts

A functional is *identifiable* if there exists a function for which the expectation is zero when the correct forecast of the functional is used as the argument. The function is termed an *identification function*. Nolde and Ziegel (2017) use these functions as the basis of calibration tests. If a consistent score is smooth with respect to the functional, its derivative is an identification function (Gneiting, 2011). An identification function for a quantile is given by:

$$V_\alpha^q\left(q_t(\alpha), y_t\right) = \alpha - I\left\{y_t \le q_t(\alpha)\right\}.$$

A forecast is *unconditionally calibrated* if the unconditional expectation of the identification function is zero, and *conditionally calibrated* if the conditional expectation is zero (Nolde and Ziegel, 2017). In this paper, the term "conditional" is synonymous with "conditional on information available at time t-1". Unconditional calibration implies that, for probability level α, the proportion of observations falling below the quantile forecasts is α. However, this is achieved by a strategic forecast set equal to unattainably high and low values for proportions α and 1-α, respectively, of the observations. Conditional calibration implies that $\Pr\left(y_t \le \hat{q}_t(\alpha)\right) = \alpha$. Christoffersen (1998) tests this by examining whether $V_\alpha^q(\hat{q}_t(\alpha), y_t)$ has zero mean and no autocorrelation. However, this is achieved for any data generating process (DGP) using Engle and Manganelli's (2004) strategic forecast of expression (2).

$$\hat{q}_t^s(\alpha) = \begin{cases} B_t & \text{if } v_t = 0 \\ A_t & \text{if } v_t = 1 \end{cases} \tag{2}$$

where A_t and B_t are values chosen to be above and below the possible values for y_t; and the v_t are outcomes of independent Bernoulli trials with probability of α for v_t=1. As $V_\alpha^q(\hat{q}_t^s(\alpha), y_t)$

has zero mean and no autocorrelation, Christoffersen's (1998) test is passed. However, $\hat{q}_t^s(\alpha)$ is not conditionally calibrated, as $V_\alpha^q(\hat{q}_t^s(\alpha), y_t)$ has non-zero conditional expectation. This is because, once $\hat{q}_t^s(\alpha)$ is known, $V_\alpha^q(\hat{q}_t^s(\alpha), y_t)$ is known. To address this, Engle and Manganelli (2004) introduce the *dynamic quantile* test, which tests whether $V_\alpha^q(\hat{q}_t(\alpha), y_t)$ has zero mean, no autocorrelation, and is independent of the forecast $\hat{q}_t(\alpha)$. For the strategic forecast of expression (2), this hypothesis is rejected because $V_\alpha^q(\hat{q}_t(\alpha), y_t)$ is not independent of $\hat{q}_t(\alpha)$.

Nolde and Ziegel (2017) consider conditional calibration in a general setting for forecasts $\hat{r}_t$ of a *k*-vector r_t of risk measures with $\mathbb{R}^k$-valued identification function $V(\hat{r}_t, y_t)$. They explain that conditional calibration requires that $E\left(V(\hat{r}_t, y_t) \mid \Psi_{t-1}\right) = 0$, where Ψ_{t-1} is the information set available at period *t*-1. They note that this is, almost surely, equivalent to the statement $E\left(h_t' V(\hat{r}_t, y_t)\right) = 0$, for all $\mathbb{R}^k$-valued h_t based on Ψ_{t-1}. Different h_t can be stacked in an $m \times k$ matrix $\mathbf{h}_t$, which they call a *test function*. This leads to the test statistic in expression (3). They draw on the results of Giacomini and White (2006) to show that this test statistic is χ^2_m-distributed asymptotically under the hypothesis of conditional calibration.

$$n\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{h}_t V\left(\hat{r}_t, y_t\right)\right)' \hat{\Omega}_n^{-1}\left(\frac{1}{n}\sum_{t=1}^{n}\mathbf{h}_t V\left(\hat{r}_t, y_t\right)\right), \tag{3}$$

where $$\hat{\Omega}_n = \frac{1}{n}\sum_{t=1}^{n}\left(\mathbf{h}_t V\left(\hat{r}_t, y_t\right)\right)\left(\mathbf{h}_t V\left(\hat{r}_t, y_t\right)\right)'.$$

This framework relies on a suitable choice for $\mathbf{h}_t$. The dynamic quantile test of Engle and Manganelli (2004) corresponds to k=1, r_t=$q_t(\alpha)$, $V(\hat{r}_t, y_t) = V_\alpha^q(\hat{q}_t(\alpha), y_t)$, and the following test function: $\boldsymbol{h}_t = \left(1,\ V_\alpha^q(\hat{q}_{t-1}(\alpha), y_{t-1}),\ \hat{q}_t(\alpha)\right)$.

3. Quantile-Bounded Interval Forecasts

In this section, for quantile-bounded interval forecasts, we review scoring functions and calibration tests, and then present a new calibration test using the framework of Nolde and Ziegel (2017). For simplicity, we consider only the symmetric case, where the interval is bounded by quantile forecasts with probability levels α and 1-α (with α<0.5). The asymmetric case is a straightforward generalisation of the results we present.

3.1. Scoring Functions for Quantile-Bounded Interval Forecasts

For a quantile-bounded interval, a consistent score is produced by summing consistent scores for the quantiles bounding the interval (Gneiting and Raftery, 2007). Consider an interval with lower bound $q_t(\alpha)$ and upper bound $q_t(1-\alpha)$. Using the quantile score of expression (1) leads to the *quantile-bounded interval score* of expression (4) (Winkler, 1972).

$$S_\alpha^{INT_q}\left(q_t(\alpha), q_t(1-\alpha), y_t\right) = \left(q_t(1-\alpha) - q_t(\alpha)\right) + \frac{1}{\alpha} I\left\{y_t \le q_t(\alpha)\right\}\left(q_t(\alpha) - y_t\right) + \frac{1}{\alpha} I\left\{y_t \ge q_t(1-\alpha)\right\}\left(y_t - q_t(1-\alpha)\right) \quad (4)$$

The score rewards narrow intervals, with observations that fall outside the interval incurring a penalty, the magnitude of which depends on the value of α (Gneiting and Raftery, 2007).

3.2. Calibration for Quantile-Bounded Interval Forecasts

An interval forecast can be said to be conditionally calibrated if the conditional probabilities of falling below $\hat{q}_t(\alpha)$ and above $\hat{q}_t(1-\alpha)$ are both equal to α. Using a Markov Chain framework, Christoffersen (1998, Section 4.2) presents a test that amounts to the null hypothesis that the variables $I(y_t \le \hat{q}_t(\alpha))$ and $I(y_t \ge \hat{q}_t(1-\alpha))$ possess no autocorrelation, no temporal cross-correlation, and both have means of α. However, for any DGP, we note that one can game the system to pass this test by drawing on the strategic quantile forecast of

expression (2) to produce the following strategic quantile-bounded interval forecast:

$$\hat{q}_t^s(\alpha)=\begin{cases} A_t & \text{if } v_t=1 \\ B_t & \text{if } v_t=2 \\ B_t & \text{if } v_t=3 \end{cases} \quad \text{and} \quad \hat{q}_t^s(1-\alpha)=\begin{cases} A_t & \text{if } v_t=1 \\ B_t & \text{if } v_t=2 \\ A_t & \text{if } v_t=3 \end{cases} \tag{5}$$

where A_t and B_t are values above and below the range of possible y_t values; and v_t is i.i.d. with categorical distribution that has probabilities of α, α and 1-2α for outcomes 1, 2 and 3, respectively. This forecast will pass Christoffersen's (1998) test because a proportion α of the observations y_t fall below $\hat{q}_t^s(\alpha)$, the same proportion fall above $\hat{q}_t^s(1-\alpha)$, and these exceedances occur with no autocorrelation, and no temporal cross-correlation.

The strategic interval forecast of expression (5) prompts us to introduce a new calibration test, which can be viewed as a synthesis of Christoffersen's (1998) test, and Engle and Manganelli's (2004) dynamic quantile test for individual quantiles. We use the framework of Nolde and Ziegel (2017), which involves the test statistic of expression (3). With k=2 and r_t=$\left(q_t(\alpha),\ q_t(1-\alpha)\right)$, the identification function is:

$$V_{\alpha,1-\alpha}^q(q_t(\alpha),q_t(1-\alpha),y_t)=\begin{pmatrix} V_\alpha^q(q_t(\alpha),y_t) \\ V_{1-\alpha}^q(q_t(1-\alpha),y_t) \end{pmatrix} \tag{6}$$

and the test function is $\boldsymbol{h}_t=\left(\boldsymbol{h}_{1t}\ \boldsymbol{h}_{2t}\right)'$, where:

$$\boldsymbol{h}_{1t}=\begin{pmatrix} 1 & V_\alpha^q(\hat{q}_{t-1}(\alpha),y_{t-1}) & V_{1-\alpha}^q(\hat{q}_{t-1}(1-\alpha),y_{t-1}) & \hat{q}_t(\alpha) \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{7}$$

$$\boldsymbol{h}_{2t}=\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & V_{1-\alpha}^q(\hat{q}_{t-1}(1-\alpha),y_{t-1}) & V_\alpha^q(\hat{q}_{t-1}(\alpha),y_{t-1}) & \hat{q}_t(1-\alpha) \end{pmatrix}. \tag{8}$$

$\boldsymbol{h}_{1t}$ enables $I\{y_t\le\hat{q}_t(\alpha)\}$ to be tested for mean of α, no autocorrelation, and no dependency on the lag of $I\{y_t\le\hat{q}_t(1-\alpha)\}$. $\boldsymbol{h}_{2t}$ enables similar testing for $I\{y_t\le\hat{q}_t(1-\alpha)\}$. Note that if the test function is chosen simply as $\boldsymbol{h}_t=\begin{pmatrix}1 & -1\end{pmatrix}$, the test examines the unconditional calibration of the interval using the hypothesis $E\left(I\{y_t\le\hat{q}_t(1-\alpha)\}-I\{y_t\le\hat{q}_t(\alpha)\}\right)=1-2\alpha$.

4. Expectile Forecasts

In this section, we review expectile scoring functions and calibration tests. We then extend calibration testing to guard against strategic prediction. An important additional aim of this section is to provide background as a basis for our more novel contribution in Section 5.

4.1. Scoring Functions for Expectile Forecasts

An expectile is the solution of an asymmetric least squares minimisation (Newey and Powell, 1987). Just as quantiles generalise the median, expectiles generalise the mean (Nolde and Ziegel, 2017). A consistent score for an expectile is the following loss function:

$$S_\tau^e\left(e_t(\tau), y_t\right) = \left|\tau - I\left\{y_t \le e_t(\tau)\right\}\right|\left(y_t - e_t(\tau)\right)^2. \tag{9}$$

where $e_t(\tau)$ is the expectile with expectile level τ, which controls the asymmetry. We term this the *expectile score*. Gneiting (2011) gives the general form of consistent score for expectiles.

4.2. Calibration for Expectile Forecasts

An identification function for an expectile is given by (Gneiting, 2011):

$$V_\tau^e\left(e_t(\tau), y_t\right) = \left|\tau - I\left\{y_t \le e_t(\tau)\right\}\right|\left(y_t - e_t(\tau)\right).$$

The expectation of this function is zero for only the true expectile, and this is the focus of calibration tests. Note that the identification function can be rewritten as

$$V_\tau^e\left(e_t(\tau), y_t\right) = \left(\tau - I\left\{y_t \le e_t(\tau)\right\}\right)\left|y_t - e_t(\tau)\right|.$$

Calibration requires that the expectation of the identification function is zero:

$$E\left(\left(\tau - I\left\{y_t \le e_t(\tau)\right\}\right)\left|y_t - e_t(\tau)\right|\right) = 0. \tag{10}$$

This can be rewritten as:

$$\frac{E\left(I\left\{y_t \le e_t(\tau)\right\}\left|y_t - e_t(\tau)\right|\right)}{E\left(\left|y_t - e_t(\tau)\right|\right)} = \tau \tag{11}$$

and
$$\frac{E\left(I\left\{y_t \le e_t(\tau)\right\}\left|y_t - e_t(\tau)\right|\right)}{E\left(I\left\{y_t > e_t(\tau)\right\}\left|y_t - e_t(\tau)\right|\right)} = \frac{\tau}{1-\tau}. \tag{12}$$

We refer to the left-hand side of expression (11) as the *expectile calibration ratio*. Newey and Powell (1987) compare expression (12) to the following analogous expression for a quantile: $F(q_t(\alpha))/(1\text{-}F(q_t(\alpha))=\alpha/(1\text{-}\alpha)$, where F_t is the distribution function. They conclude that "expectiles are determined by tail expectations in the same way that quantiles are determined by the distribution". Expression (12) can be rewritten as expression (13).[2] Fig. 1 illustrates this expression, showing that the expectile marks the point on the distribution function F_t for which the ratio of area Z_1 to area Z_2 equals $\tau/(1\text{-}\tau)$.

$$\frac{\int_{-\infty}^{e_t(\tau)} F_t(y)dy}{\int_{e_t(\tau)}^{\infty}\left(1-F_t(y)\right)dy} = \frac{\tau}{1-\tau} \tag{13}$$

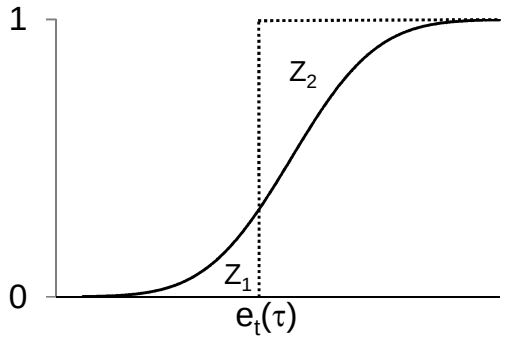


Fig. 1. For distribution F_t, an expectile $e_t(\tau)$ is such that the ratio of area Z_1 to area Z_2 equals $\tau/(1\text{-}\tau)$.

Nolde and Ziegel (2017) describe a calibration test for y_t expressed as $y_t=\mu_t+\sigma_t z_t$, where μ_t is the mean, σ_t is the standard deviation, and z_t is a sequence of i.i.d. random variables. The expectile forecast of z_t is $\hat{e}_{z,t}(\tau) = (\hat{e}_t(\tau) - \hat{\mu}_t)/\hat{\sigma}_t$, where $\hat{\mu}_t$ and $\hat{\sigma}_t$ are mean and standard deviation forecasts. Nolde and Ziegel (2017) propose that $V_\tau^e(\hat{e}_{z,t}(\tau), z_t)$ is tested for zero mean, but it is natural also to test that it is i.i.d. However, for any DGP, the following strategic expectile forecast for y_t satisfies both these conditions, even though it is a poor forecast:

[2] Expression (12) is the basis on which Taylor (2008) uses expectiles, in a non-standard way, to obtain estimates of value at risk and expected shortfall (see also Gerlach and Wang, 2020).

$$\hat{e}_t^s(\tau) = \begin{cases} \hat{\mu}_t + B\hat{\sigma}_t & \text{if } \nu_t = 0 \\ \hat{\mu}_t + A\hat{\sigma}_t & \text{if } \nu_t = 1 \end{cases}$$

which can also be expressed as a strategic expectile forecast for z_t:

$$\hat{e}_{z,t}^s(\tau) = \begin{cases} B & \text{if } \nu_t = 0 \\ A & \text{if } \nu_t = 1 \end{cases} \tag{14}$$

where ν_t is the outcome of a sequence of independent Bernoulli trials with probability of α for ν_t=1; and A and B are values above and below the possible values for z_t satisfying the following:

$$B = -\frac{\theta}{(1-\theta)}\frac{(1-\tau)}{\tau}A. \tag{15}$$

θ can be any chosen value between 0 and 1. For example, if $\theta=\tau$, we get B=-A, or if θ=0.5, we get B=-(1-τ)A/τ. Expression (15) implies that $V_\tau^e(\hat{e}_{z,t}^s(\tau), z_t)$ has zero mean, and $V_\tau^e(\hat{e}_{z,t}^s(\tau), z_t)$ has no autocorrelation because z_t falls below $\hat{e}_{z,t}^s(\tau)$ randomly. Therefore, a calibration test should not simply test for zero mean and no autocorrelation in the identification function.

To address the problem of the strategic forecast, we propose a conditional calibration test similar in spirit to Engle and Manganelli's (2004) dynamic quantile test. We express this new test using the framework of Nolde and Ziegel (2017), discussed in relation to expression (3). We set k=1, $r_t = e_{z,t}(\tau)$ and $V(\hat{r}_t, z_t) = V_\tau^e(\hat{e}_{z,t}(\tau), z_t)$, and use the following test function:

$$\boldsymbol{h}_t = \left(1, \quad V_\tau^e(\hat{e}_{z,t-1}(\tau), z_{t-1}), \quad \hat{e}_{z,t}(\tau)\right).$$

The first entry in $\boldsymbol{h}_t$ corresponds to the calibration test of Nolde and Ziegel (2017), which has $E\left(V_\tau^e(\hat{e}_{z,t}(\tau), z_t)\right) = 0$ as the hypothesis. The second entry in $\boldsymbol{h}_t$ enables the testing for autocorrelation in the identification function $V_\tau^e(\hat{e}_{z,t}(\tau), z_t)$, and the third entry guards against the strategic forecast of expression (14). We call this the *dynamic expectile* test.

5. Expectile-Bounded Interval Forecasts

The most common way to express forecast uncertainty is a quantile-bounded interval. In this section, as an alternative, we introduce expectile-bounded intervals. We consider symmetric expectile bounds with expectile levels τ and 1-τ, where τ<0.5. We first informally compare quantile-bounded and expectile-bounded intervals. We then introduce a scoring function and calibration test for this new form of interval forecast.

5.1. An Informal Comparison of Quantile-Bounded and Expectile-Bounded Intervals

In Section 1, we discussed the appeal of using expectile-bounded intervals. A fundamental motivation is that, relative to expectiles, quantiles have the disadvantage of not conveying information regarding the size of potential exceedances beyond the quantile. To expand on this, we now briefly compare quantiles and expectiles. Yao and Tong (1996) show that the quantile $q_t(\alpha)$ is equal to the expectile $e_t(\tau)$ if τ is given by the following expression:

$$\tau(\alpha) = \frac{\alpha q_t(\alpha) - \int_{-\infty}^{q_t(\alpha)} y dF_t(y)}{E(y_t) - 2\int_{-\infty}^{q_t(\alpha)} y dF_t(y) - (1-2\alpha)q_t(\alpha)} \tag{16}$$

Following Kuan et al. (2009), Fig. 2 plots the function $\tau(\alpha)$ for the following distributions: normal, Student t distribution with 4 degrees of freedom, and the skewed t distribution described by Christoffersen (2012, Chapter 6, Section 7), with parameters d_1=4 and d_2=0.3. For the normal and Student t distributions, the curves are below the 45^0 line for α<0.5, and above the line for α>0.5, while for the skewed t distribution, the curve lies mostly below the 45^0 line. For a normal distribution, expression (16) and Fig. 2 indicate that the interval bounded by $q_t(\alpha)$ and $q_t(1\text{-}\alpha)$ with α=0.01 is equal to the interval bounded by $e_t(\tau)$ and $e_t(1\text{-}\tau)$ with τ=0.00145; the quantile-bounded interval corresponding to α=0.025 is equal to the expectile-bounded interval corresponding to τ=0.00477; and the quantile-bounded interval for α=0.10 is equal to the expectile-bounded interval for τ=0.03438.

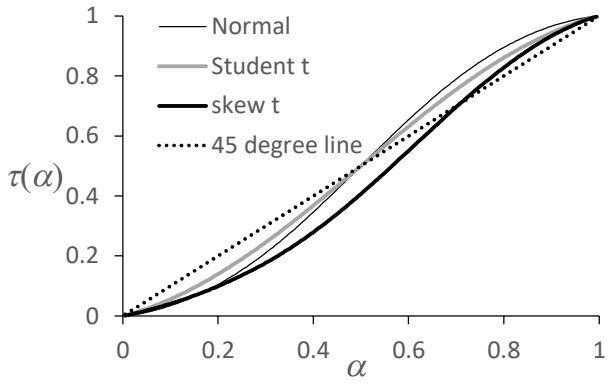


Fig. 2. Empirical illustration of the function $\tau(\alpha)$ of expression (16). $\tau(\alpha)$ is the expectile level for which the expectile $e_t(\tau(\alpha))$ is equal to the quantile $q_t(\alpha)$.

Table 1
Bounds on three quantile-bounded intervals for distributions with zero mean and unit variance.

	Normal	Student t	Skewed t
$q_t(0.01)$	-2.33	-2.65	-1.99
$q_t(0.99)$	2.33	2.65	3.16
$q_t(0.025)$	-1.96	-1.96	-1.56
$q_t(0.975)$	1.96	1.96	2.27
$q_t(0.10)$	-1.28	-1.08	-0.99
$q_t(0.90)$	1.28	1.08	1.14

Table 2
Bounds on three expectile-bounded intervals for distributions with zero mean and unit variance.

	Normal	Student t	Skewed t
$e_t(0.00145)$	-2.33	-3.48	-2.42
$e_t(0.99855)$	2.33	3.48	4.36
$e_t(0.00477)$	-1.96	-2.50	-1.81
$e_t(0.99523)$	1.96	2.50	3.07
$e_t(0.03438)$	-1.28	-1.32	-1.05
$e_t(0.96562)$	1.28	1.32	1.54

Table 1 presents three quantile-bounded intervals for the same three distributions: normal, Student t and skewed t. Note that we have set all three to have zero mean and unit variance. Table 2 provides intervals bounded by the expectiles $e_t(\tau)$ and $e_t(1\text{-}\tau)$ with τ=0.00145, 0.00477 and 0.03438, which are the values of τ that we identified in the previous paragraph. For the normal distribution, the first column of values in Table 2 confirms that these expectile-

bounded intervals are identical to the quantile-bounded intervals in the first column of values in Table 1. For the Student t distribution, comparing the second columns of values in Tables 1 and 2, we see that moving from the normal distribution to this fat tailed distribution has a greater impact on the widening of the expectile-bounded intervals than the quantile-bounded intervals. For the skewed t distribution, comparing the tables shows that moving from the symmetric distributions to this skewed distribution has a greater impact on the asymmetry of the expectile-bounded interval than the quantile-bounded interval.

A practical implication of the numerical illustrations in this section is that, if the kurtosis or skewness in a distribution change over time, it is likely to be more apparent from an expectile-bounded interval than a quantile-bounded interval.

5.2. Scoring Functions for Expectile-Bounded Interval Forecasts

By analogy with the quantile-bounded interval, a consistent score for an expectile-bounded interval can be produced by summing consistent scores for the expectiles, $e_t(\tau)$ and $e_t(1-\tau)$, bounding the interval. Using the expectile score of expression (9) leads to the following new *expectile-bounded interval score*:

$$\begin{aligned} S_{\tau}^{INT_e}\left(e_t(\tau), e_t(1-\tau), y_t\right) = &\left(y_t - e_t(\tau)\right)^2 + \left(y_t - e_t(1-\tau)\right)^2 \\ &+ \frac{1}{\tau} I\left\{y_t \le e_t(\tau)\right\}(1-2\tau)\left(y_t - e_t(\tau)\right)^2 \\ &+ \frac{1}{\tau} I\left\{y_t \ge e_t(1-\tau)\right\}(1-2\tau)\left(y_t - e_t(1-\tau)\right)^2 . \end{aligned} \tag{17}$$

The last two terms in this expression impose penalties when an observation falls outside the interval. If these terms were removed, the score would be minimised by setting both bounds to be the mean of y_t. The score, therefore, rewards narrow intervals subject to penalties for interval exceedance, which was also the case for the quantile-bounded interval score of expression (4).

5.3. Calibration for Expectile-Bounded Interval Forecasts

Using expression (10) for each interval bound, $e_t(\tau)$ and $e_t(1-\tau)$, we have the following conditions for calibration at each bound:

$$E\left(\left(\tau - I\left\{y_t \le e_t(\tau)\right\}\right)\left|y_t - e_t(\tau)\right|\right) = 0\,, \tag{18}$$

$$E\left(\left(1-\tau - I\left\{y_t \le e_t(1-\tau)\right\}\right)\left|y_t - e_t(1-\tau)\right|\right) = 0\,. \tag{19}$$

Subtracting expression (19) from expression (18) leads to expression (20). We term the left-hand side of this expression the *expectile-bounded interval calibration ratio*.

$$\frac{E\left(I\left\{y_t \le e_t(\tau)\right\}\left|y_t - e_t(\tau)\right| + I\left\{y_t \ge e_t(1-\tau)\right\}\left|y_t - e_t(1-\tau)\right|\right)}{\left(e_t(1-\tau) - e_t(\tau)\right)/2} = \frac{2\tau}{1-2\tau} \tag{20}$$

This expression provides insight and helps with the interpretability of an expectile-bounded interval. It is the ratio of the expectation of the size of interval exceedances to half the interval width, which is the average distance of a point within the interval to the interval bounds. As the ratio equals $2\tau/(1-2\tau)$, we view expression (20) as the analogue of an expression for a quantile-bounded interval involving a ratio of the expected number of observations outside the interval to the expected number within.

Rewriting expression (20) in terms of the distribution function F_t gives expression (21), which we feel also helps provide insight and interpretability for expectile-bounded intervals. Fig. 3 represents this expression graphically, showing that the expectile bounds are such that $2\tau/(1-2\tau)$ is equal to the ratio of the sum of areas Z_1 and Z_2 to half the interval width.

$$\frac{\int_{-\infty}^{e_t(\tau)} F_t(y)dy + \int_{e_t(1-\tau)}^{\infty}\left(1-F_t(y)\right)dy}{\left(e_t\left(1-\tau\right) - e_t\left(\tau\right)\right)/2} = \frac{2\tau}{1-2\tau} \tag{21}$$

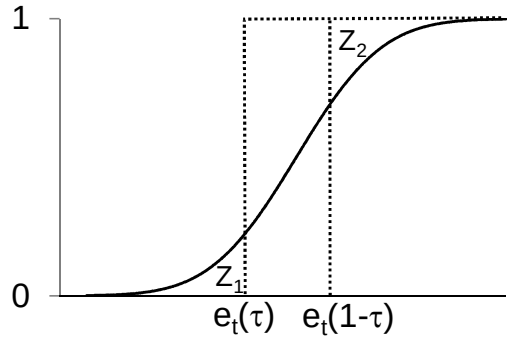


Fig. 3. For distribution F_t, the interval bounded by expectiles $e_t(\tau)$ and $e_t(1\text{-}\tau)$ is such that the ratio of the sum of areas Z_2 and Z_1 to half the interval width is equal to $2\tau/(1\text{-}2\tau)$.

We now introduce a conditional calibration test for expectile-bounded interval forecasts. The test draws on the calibration test for quantile-bounded interval forecasts of Section 3.2, and the dynamic expectile calibration test of Section 4.2. We again use the framework of Nolde and Ziegel (2017), which involves the test statistic of expression (3). With k=2 and $r_t=\left(\hat{e}_{zt}(\tau),\ \hat{e}_{zt}(1-\tau)\right)$, the identification function is:

$$V^e_{\tau,1-\tau}(\hat{e}_{zt}(\tau),\ \hat{e}_{zt}(1-\tau), z_t) = \begin{pmatrix} V^e_{\tau}(\hat{e}_{zt}(\tau), z_t) \\ V^e_{1-\tau}(\hat{e}_{zt}(1-\tau), z_t) \end{pmatrix} \tag{22}$$

and the test function is $\boldsymbol{h}_t = \left(\boldsymbol{h}_{1t}\ \ \boldsymbol{h}_{2t}\right)'$, where:

$$\boldsymbol{h}_{1t} = \begin{pmatrix} 1 & V^e_{\tau}(\hat{e}_{zt-1}(\tau), z_{t-1}) & V^e_{1-\tau}(\hat{e}_{zt-1}(1-\tau), z_{t-1}) & \hat{e}_{zt}(\tau) \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{23}$$

$$\boldsymbol{h}_{2t} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & V^e_{1-\tau}(\hat{e}_{zt-1}(1-\tau), z_{t-1}) & V^e_{\tau}(\hat{e}_{zt-1}(\tau), z_{t-1}) & \hat{e}_{zt}(1-\tau) \end{pmatrix}. \tag{24}$$

Note that if the test function is chosen simply as $\boldsymbol{h}_t = \begin{pmatrix} 1 & -1 \end{pmatrix}$, the resultant test examines the unconditional calibration of the interval with expression (20) as the null hypothesis.

6. Empirical Illustration of Interval Forecasts and Their Evaluation

We now use electricity price data to illustrate the ideas of Sections 3 and 5 regarding interval forecasts. Electricity prices exhibit time-variation in both the mean and volatility, with a common feature being the presence of sizeable spikes. Nowotarksi and Weron (2018)

describe how interest is rapidly growing in the use of probabilistic forecasting of electricity prices, with (quantile-bounded) interval forecasts commonly used to summarise risk.

6.1. Electricity Price Data

We used hourly Nord Pool market clearing prices for 2013 to 2018, inclusive, downloaded from the Nordic power exchange website (www.nordpoolgroup.com/). Each day, the price is set for each hour of the following day. In view of this, we follow Weron and Misiorek (2008) by forecasting the Nord Pool price separately for each hour of the day using historical data for only that hour. For each hour, we used a three-year rolling window, each consisting of 3×52×7=1092 daily observations, for repeated re-estimation of method parameters. This led to day-ahead out-of-sample forecasts for the final 1255 days. The daily series of prices for the twelfth hour of the day is plotted in Fig. 4. The plot shows that the series has time-varying mean and variance, and is positively skewed.

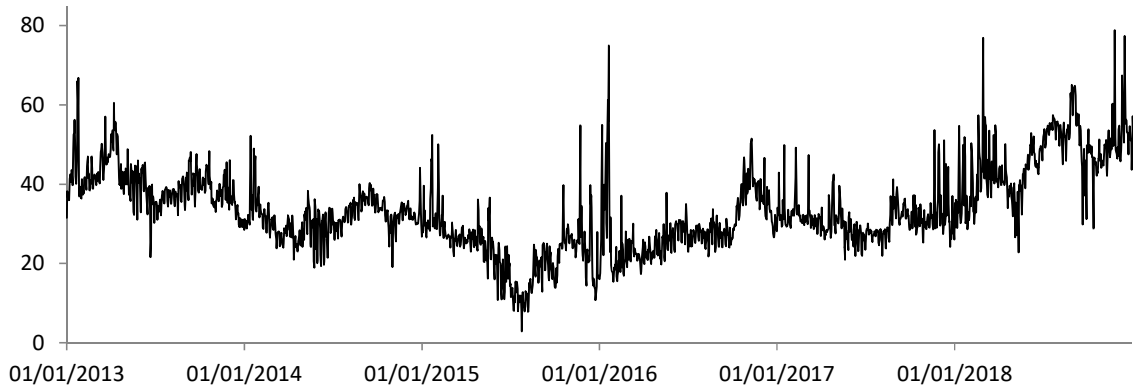


Fig. 4. Daily Nord pool electricity prices (EUR/MWh) for the twelfth hour of the day.

6.2. Probabilistic Forecasting Methods

Many different methods have been considered for the probabilistic forecasting of day-ahead electricity prices (Nowotarski and Weron, 2018). In this paper, for our illustrative purposes, we implement relatively simple approaches, based on the model of expression (25),

which Weron and Misiorek (2008) find to be competitive for Nord pool prices. A similar model is considered by Nowotarski and Weron (2018).

$$p_{h,t} = \phi_{h,0} + \phi_{h,1} p_{h,t-1} + \phi_{h,2} p_{h,t-2} + \phi_{h,3} p_{h,t-7} + \phi_{h,4} mp_{t-1} + d_{h,1} D_{Mon,t} + d_{h,2} D_{Sat,t} + d_{h,3} D_{Sun,t} + \varepsilon_{h,t} \quad (25)$$

where $p_{h,t}$ is the log price for hour *h* on day *t*; mp_t is the minimum value of the log price on day *t*; $D_{Mon,t}$, $D_{Sat,t}$ and $D_{Sun,t}$ are binary variables indicating whether day *t* is a Monday, Saturday or Sunday, respectively; the $\phi_{h,i}$ and $d_{h,i}$ are parameters; and, for each *h*, the $\varepsilon_{h,t}$ are independent and Gaussian, with zero mean and constant variance. Our use of the log transformation was to stabilise the variance (see Uniejewski et al., 2017). We estimated the model using least squares. Weron and Misiorek (2008) consider the inclusion of a temperature variable as a proxy for temperature forecasts, but it does not lead to improved forecast accuracy, and so, for simplicity, we did not consider this in our analysis.

We produced probabilistic forecasts from the model of expression (25) using four different approaches. The first two were considered by Weron and Misiorek (2008). The first used a Gaussian assumption for the log prices, and we term this the *AR-N* method. The second approach produced a distributional forecast for the log price using the empirical distribution of the residuals from the model, centred at the model's forecast of the mean. We refer to this as the *AR-Emp* method. We also estimated expression (25) with $\varepsilon_{h,t}$ specified as the skewed t distribution described by Christoffersen (2012, Chapter 6, Section 7), with variance modelled using the asymmetric GJR-GARCH model of Glosten et al. (1993). We refer to this third approach as the *AR-GJR-GARCH-SkewT* method. Asymmetric GARCH models have been considered previously for electricity prices (see, for example, Knittel and Roberts, 2005). We also produced distributional forecasts using the empirical distribution of the standardised residuals. We refer to this fourth approach as the *AR-GJR-GARCH-Emp* method.

The four methods that we have described produce distributional forecasts for log price. For the *AR-N* and *AR-GJR-GARCH-SkewT* methods, we simulated 10^4 values from the log

price distribution, and applied the exponential transformation to generate simulated values from which we produced distributional, quantile and expectile forecasts for price. To estimate an expectile from the 10^4 simulated values, we used iteratively reweighted least squares to minimise the expectile score of expression (9) summed over the estimation sample (see Newey and Powell, 1987). For the *AR-Emp* and *AR-GJR-GARCH-Emp* methods, we followed the same procedure based on the empirically constructed sample of log price values.

6.3. Quantile-Bounded and Expectile-Bounded Interval Forecasts

To illustrate each type of interval forecast, quantile-bounded and expectile-bounded, we felt it would be useful to consider two different intervals. For quantile-bounded intervals, we estimated intervals bounded by the quantiles $q_t(0.025)$ and $q_t(0.975)$, and narrower intervals bounded by $q_t(0.10)$ and $q_t(0.90)$. For expectile-bounded intervals, we chose to consider two intervals of similar widths to the quantile-bounded intervals. To achieve this, we experimented with different choices of the expectile level τ in order to find expectile-bounded intervals that contained approximately 95% and 80% of the observations. This led us to consider intervals bounded by $e_t(0.01)$ and $e_t(0.99)$, and narrower intervals bounded by $e_t(0.05)$ and $e_t(0.95)$.

For the twelfth hour of the day, and the first half of 2018, Fig. 5 shows intervals bounded by forecasts of $q_t(0.025)$ and $q_t(0.975)$, and by forecasts of $e_t(0.01)$ and $e_t(0.99)$, where all forecasts were out-of-sample predictions from the AR-GJR-GARCH-SkewT method. The plot shows that, for each day, the lower bounds of the two intervals are similar, but the upper bounds differ, with the upper expectile bound often more extreme than the upper quantile bound. This difference can be explained by there being more frequent extremes in the upper tail, and by the tendency for extremes to have a greater impact on expectiles than quantiles. This relates to our comment at the end of Section 5.1 that time-varying kurtosis and skewness are likely to be more apparent from an expectile-bounded interval than a quantile-bounded interval.

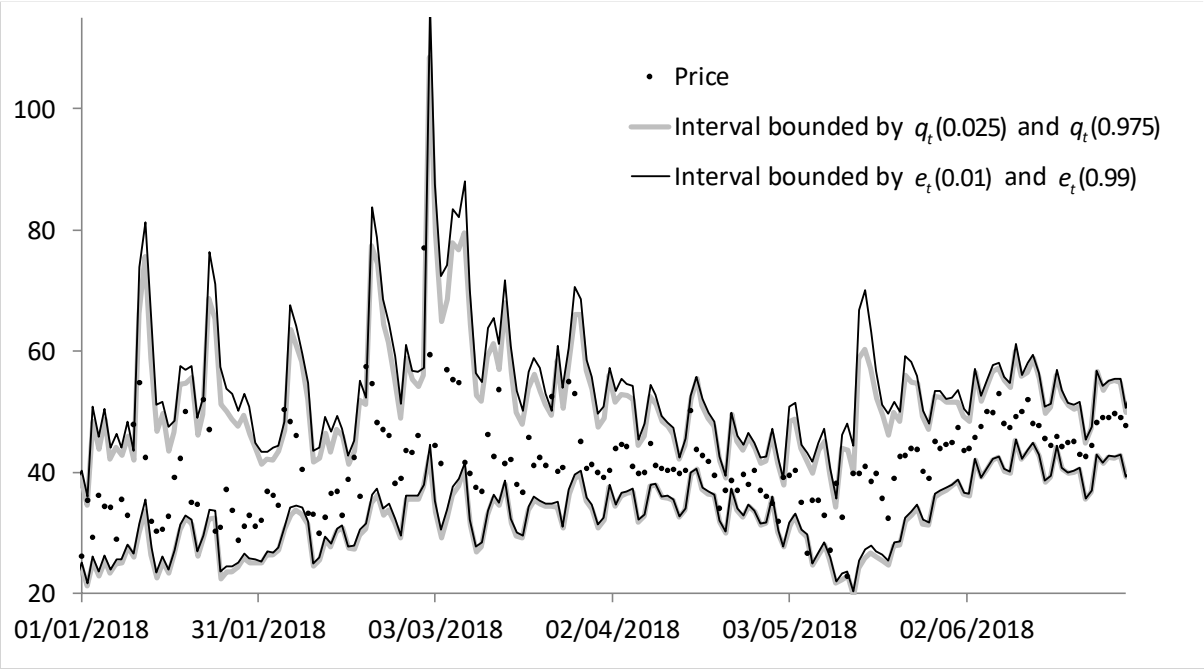


Fig. 5. Daily prices (EUR/MWh) for the twelfth hour of the day with intervals bounded by out-of-sample forecasts of $q_t(0.025)$ and $q_t(0.975)$, and by out-of-sample forecasts of $e_t(0.01)$ and $e_t(0.99)$. Both sets of interval forecasts were produced by the AR-GJR-GARCH-SkewT method.

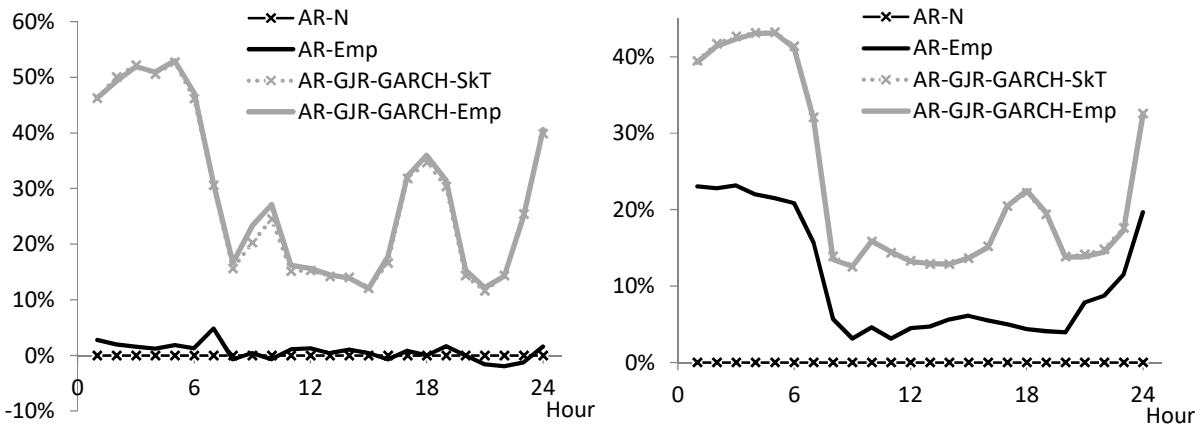
6.4. Evaluation of Out-of-Sample Quantile-Bounded Interval Forecasts

For each hour of the day, Fig. 6 summarises the results for the quantile-bounded interval score of expression (4). The figure presents skill scores, which were computed as 1 minus the ratio of a method's mean score to that of the AR-N method. Higher values are preferable. The AR-GJR-GARCH methods clearly outperform the AR methods for both the wider and narrower intervals.

Fig. 7 summarises unconditional calibration of the intervals, assessed by the percentage of the out-of-sample periods falling within the interval for each hour of the day. The ideal value is indicated in each of the two plots by the horizontal dashed line. With this in mind, the best results correspond to AR-GJR-GARCH-Emp.

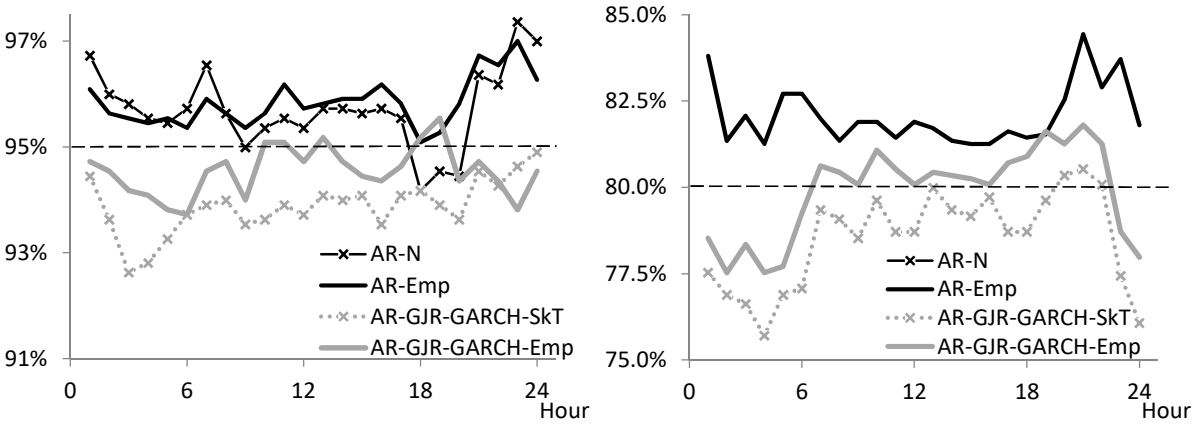
Fig. 8 presents p-values for the conditional calibration test based on expressions (6) to (8) and the framework of Nolde and Ziegel (2017). The p-values are generally rather small, indicating many cases of rejection of the hypothesis of conditional calibration. (This was consistent with our finding that conditional calibration was rejected for either or both of the

quantile forecasts bounding the interval when evaluated using the dynamic quantile test.)



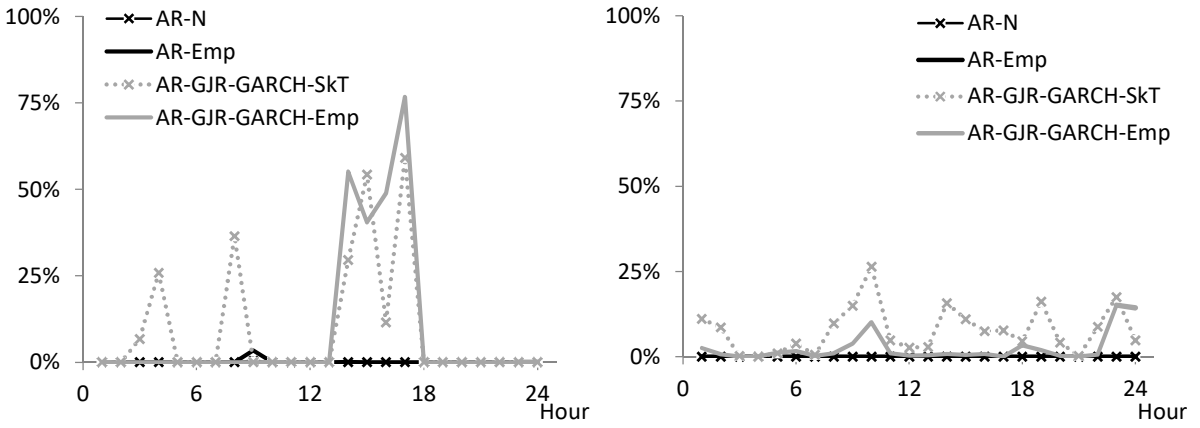
(a) $q_t(0.025)$ and $q_t(0.975)$ (b) $q_t(0.10)$ and $q_t(0.90)$

Fig. 6. Skill scores for quantile-bounded intervals. Higher values are better.



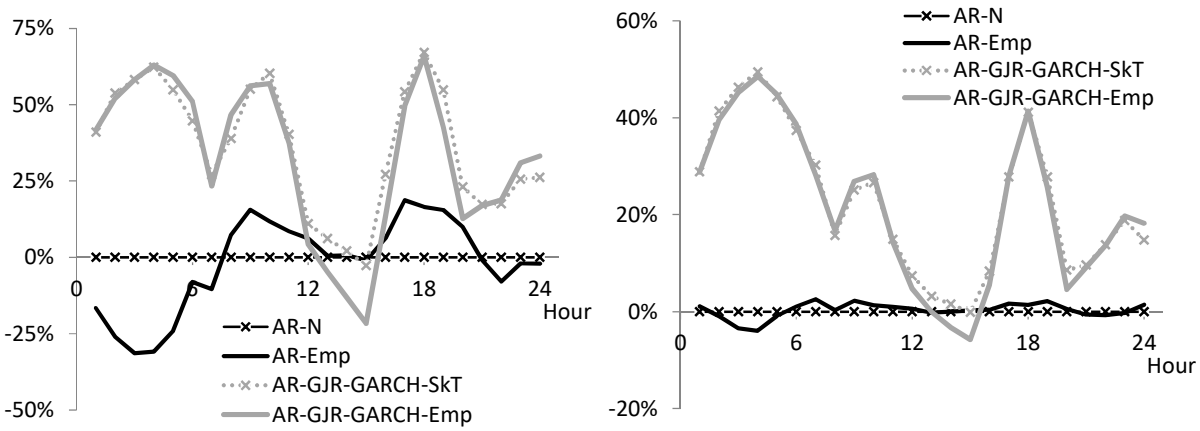
(a) $q_t(0.025)$ and $q_t(0.975)$ (b) $q_t(0.10)$ and $q_t(0.90)$

Fig. 7. Unconditional calibration for quantile-bounded intervals assessed using interval coverage percentage. Ideal value is indicated by horizontal dashed line.



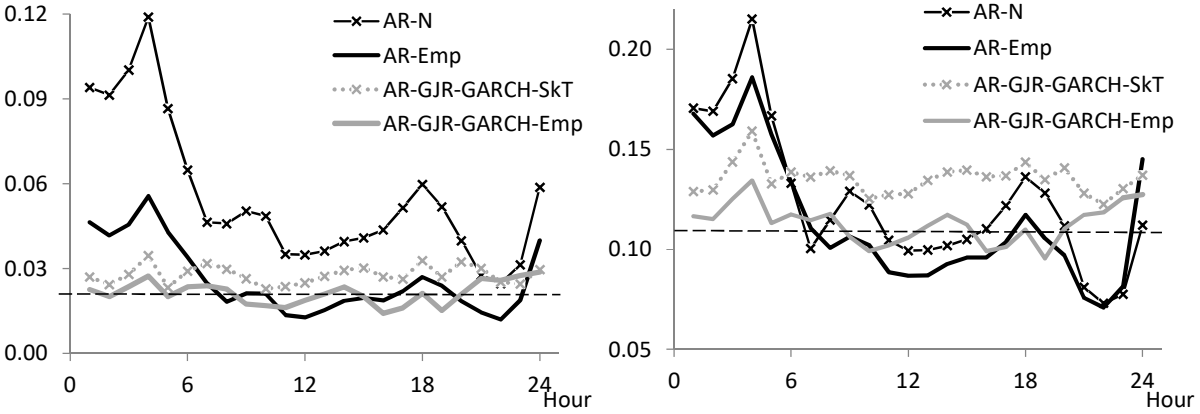
(a) $q_t(0.025)$ and $q_t(0.975)$ (b) $q_t(0.10)$ and $q_t(0.90)$

Fig. 8. Conditional calibration test p-values for quantile-bounded intervals. Higher values are better, with low values indicate rejection of conditional calibration.



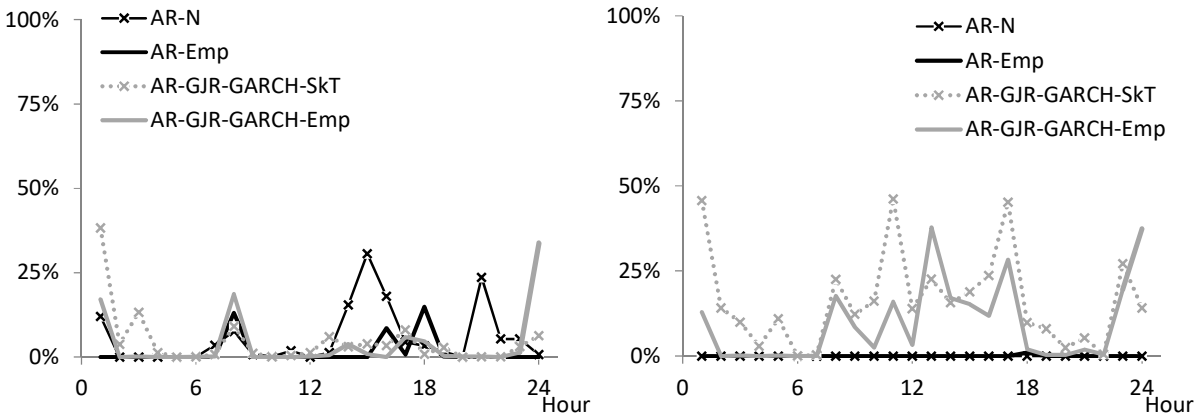
(a) $e_t(0.01)$ and $e_t(0.99)$ (b) $e_t(0.05)$ and $e_t(0.95)$

Fig. 9. Skill scores for expectile-bounded intervals. Higher values are better.



(a) $e_t(0.01)$ and $e_t(0.99)$ (b) $e_t(0.05)$ and $e_t(0.95)$

Fig. 10. Unconditional calibration for expectile-bounded intervals assessed using expectile-bounded interval calibration ratio of expression (20). Ideal value is indicated by horizontal dashed line.



(a) $e_t(0.01)$ and $e_t(0.99)$ (b) $e_t(0.05)$ and $e_t(0.95)$

Fig. 11. Conditional calibration test p-values for expectile-bounded intervals. Higher values are better, with low values indicate rejection of conditional calibration.

6.5. Evaluation of Out-of-Sample Expectile-Bounded Interval Forecasts

Fig. 9 presents the skill score results for the expectile-bounded interval score of expression (17). The AR-GJR-GARCH methods outperform the AR methods. It is interesting to see that the relative performances of the four methods in Fig. 9 differ from the corresponding results for the quantile-bounded interval in Fig. 6. In contrast with Fig. 6, the results differ slightly for the two AR-GJR-GARCH methods in Fig. 9, with AR-GJR-GARCH-SkT performing better in terms of the expectile-bounded interval score.

As a summary of unconditional calibration, Fig. 10 presents the sample estimate of the expectile-bounded interval calibration ratio of expression (20) computed from the out-of-sample observations. The ideal value is $2\tau/(1-2\tau)$, and this is indicated in the plots of Fig. 10 by the horizontal dashed lines. For the narrower interval (Fig. 10(b)), the ideal value is 0.111, which, as we discussed in Section 5.3, implies that the average size of interval exceedances should be about 11% of the average distance of a point within the interval to the interval bounds. The best results correspond to AR-GJR-GARCH-Emp, which is also the case with the unconditional calibration test results for the quantile-bounded intervals in Fig. 7.

Fig. 11 presents p-values for the conditional calibration test based on expressions (22) to (24) and the framework of Nolde and Ziegel (2017). Overall, the p-values are larger than for the corresponding test for the quantile-bounded intervals in Fig. 8. In Fig. 11, for the wider interval (Fig. 11(a)), it is difficult to rank the methods, but for the narrower interval (Fig. 11(b)), the results are better for the two AR-GJR-GARCH methods.

7. Simulation Study

We used simulated data to provide a brief check on the measures and tests that we have considered in the paper. As our work draws heavily on the framework of Nolde and Ziegel (2017), we implemented a similar simulation study to the one that they present. This involved data generated from the following AR(1)-GARCH(1,1) process:

$$y_t = -0.05 + 0.3y_{t-1} + \varepsilon_t \qquad \varepsilon_t = \sigma_t Z_t \qquad \sigma_t^2 = 0.01 + 0.1\varepsilon_{t-1}^2 + 0.85\sigma_{t-1}^2 \tag{26}$$

where the Z_t form a sequence of independent random terms generated from a skewed t distribution of the type described by Christoffersen (2012, Chapter 6, Section 7), with parameters d_1=4 and d_2=0.3. These two parameters differ from those used by Nolde and Ziegel (2017), because they used a different formulation for the skewed t distribution. We followed Nolde and Ziegel (2017) in using a series of 5500 observations, with rolling windows of 500 periods used for repeated re-estimation of parameters, and the final 5000 periods used to compare out-of-sample forecast accuracy. We implemented the same four individual methods that we considered for the electricity price data. As a benchmark, we also produced forecasts from the true data generating process, which we refer to as *True DGP*. Out-of-sample results for the quantile-bounded and expectile bounded interval forecasts are presented in Tables 3 and 4, respectively.

The rankings of methods in the two tables are similar for each of the three forms of evaluation. In both tables, the skill score results are comfortably better for the methods that employ AR-GARCH. As AR-GARCH-SkT and AR-GARCH-Emp use the correct choice of AR-GARCH model, it is not surprising to see them perform well. Indeed, although the true DGP has the highest skill score for the narrower intervals in Tables 3 and 4, it is actually slightly outperformed for the wider intervals. The conditional calibration test results are better in both tables for the methods that are based on AR-GARCH. In terms of unconditional calibration, the results for the True DGP are not particularly strong, with AR-Emp actually performing better overall. However, the skill score and conditional calibration results indicate that AR-Emp is only capturing the unconditional characteristics of the simulated data, but that it fails to time-varying features.

Table 3
For simulated data, quantile-bounded interval forecasts evaluated using the skill score, and unconditional and conditional calibration.

Quantiles bounding interval	$q_t(0.025)$ & $q_t(0.975)$	$q_t(0.10)$ & $q_t(0.90)$
Skill scores		
AR-N	0.0	0.0
AR-Emp	2.3	2.6
AR-GARCH-SkT	11.5	7.8
AR-GARCH-Emp	11.1	7.7
True DGP	11.2	8.3
Unconditional calibration assessed using interval coverage percentage		
AR-N	94.2	86.3
AR-Emp	94.3	79.7
AR-GARCH-SkT	94.6	79.9
AR-GARCH-Emp	94.5	79.8
True DGP	94.1	78.9
Conditional calibration test p-values for quantile-bounded intervals		
AR-N	0.00	0.00
AR-Emp	0.00	0.00
AR-GARCH-SkT	0.39	0.29
AR-GARCH-Emp	0.28	0.05
True DGP	0.26	0.36

Notes: Higher skill scores are better. For unconditional calibration, ideal is 95% for interval bounded by $q_t(0.025)$ and $q_t(0.975)$, and 80% for interval bounded by $q_t(0.10)$ and $q_t(0.90)$. For conditional calibration, higher p-values are better.

Table 4
For simulated data, expectile-bounded interval forecasts evaluated using the skill score, and unconditional and conditional calibration.

Expectiles bounding interval	$e_t(0.01)$ & $e_t(0.99)$	$e_t(0.05)$ & $e_t(0.95)$
Skill scores		
AR-N	0.0	0.0
AR-Emp	8.3	1.3
AR-GARCH-SkT	19.5	9.1
AR-GARCH-Emp	19.7	8.7
True DGP	19.5	9.3
Unconditional calibration assessed using expectile-bounded interval calibration ratio		
AR-N	0.042	0.122
AR-Emp	0.023	0.117
AR-GARCH-SkT	0.020	0.114
AR-GARCH-Emp	0.023	0.115
True DGP	0.023	0.125
Conditional calibration test p-values for expectile-bounded intervals		
AR-N	0.00	0.00
AR-Emp	0.05	0.00
AR-GARCH-SkT	0.00	0.31
AR-GARCH-Emp	0.00	0.20
True DGP	0.06	0.31

Notes: Higher skill scores are better. For unconditional calibration, ideal is $2\tau/(1-2\tau)$, which, for interval bounded by $e_t(0.01)$ and $e_t(0.99)$ is 0.020, and for interval bounded by $e_t(0.05)$ and $e_t(0.95)$ is 0.111. For conditional calibration, higher p-values are better.

8. Concluding Comments

In this paper, we have provided a review of scoring functions and calibration tests for quantiles, expectiles and interval forecasts. Using the framework of Nolde and Ziegel (2017), we have presented new conditional calibration tests for quantile-bounded interval forecasts and expectile forecasts, which include elements to guard against strategic forecasting. In view of the usefulness of interval forecasts, and the advantages of expectiles, we propose expectile-bounded intervals. We present a scoring function, a calibration test, and an interpretation for this new type of interval forecast. We note that a broader generalisation of quantile-bounded intervals can be provided by using, as interval bounds, the M-quantiles of Breckling and

Chambers (1988). Expectile-bounded intervals are a special case of this.

Acknowledgements

We would like to thank the Editor-in-Chief and two referees for providing very useful comments on the paper.

References

Artzner, P., Delbaen, F., Eber, J.M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance, 9*(3) 203-228.

Breckling, J., & Chambers, R. (1988). M-quantiles. *Biometrika, 75*(4) 761-771.

Christoffersen, P.F. (1998). Evaluating interval forecasts. *International Economic Review, 39*(4) 841-862.

Christoffersen, P.F. (2012). *Elements of Financial Risk Management*. 2nd edition, Waltham: Academic Press

Engle, R.F., & Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics, 22*(4) 367-381.

Gerlach, R., & Wang, C. (2020). Semi-parametric dynamic asymmetric Laplace models for tail risk forecasting, incorporating realized measures. *International Journal of Forecasting*, forthcoming.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica, 74*(6) 1545-1578.

Glosten, L.R., Jagannathan, R., & 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*(5) 1779-1801.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association, 106*(494) 746-762.

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and its Application*, *1*(1) 125-151.

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

Knittel, C.R., & Roberts, M.R. (2005). An empirical examination of restructured electricity prices. *Energy Economics*, *27*(5) 791–817.

Koenker, R.W. (2005). *Quantile Regression.* Cambridge, UK: Cambridge University Press.

Koenker R., Machado J.A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, *94*(448)1296-1310.

Kuan, C. M., Yeh, J. H., & Hsu, Y. C. (2009). Assessing value at risk with care, the conditional autoregressive expectile models. *Journal of Econometrics*, *150*(2), 261-270.

Lichtendahl Jr, K.C., Grushka-Cockayne, Y. and Pfeifer, P.E. (2013). The wisdom of competitive crowds. *Operations Research*, *61*(6) 1383-1398.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*(4) 802-808.

Newey, W.K., & Powell, J.L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, *55*(4) 819-847.

Nolde, N., & Ziegel, J.F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Annals of Applied Statistics*, *11*(4) 1833-1874.

Nowotarski, J., & Weron, R. (2018). Recent advances in electricity price forecasting: A review of probabilistic forecasting. *Renewable and Sustainable Energy Reviews*, *81*(8) 1548-1568.

Olszewski, W. (2015). Calibration and expert testing. In *Handbook of Game Theory with Economic Applications*, edited by H. P. Young and S. Zamir, *4* 949-984, Elsevier.

Taylor J.W. (1999). Evaluating volatility and interval forecasts. *Journal of Forecasting*, *18*(2) 111-128.

Taylor, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. *Journal*

of Financial Econometrics, *6*(2) 231-252.

Taylor, J. W. (2020). A strategic predictive distribution for tests of probabilistic calibration. *International Journal of Forecasting*, forthcoming.

Uniejewski, B., Weron, R., & Ziel, F. (2017). Variance stabilizing transformations for electricity spot price forecasting. *IEEE Transactions on Power Systems*, *33*(2) 2219-2229.

Weron, R., & Misiorek, A. (2008). Forecasting spot electricity prices: a comparison of parametric and semiparametric time series models. *International Journal of Forecasting*, *24*(4) 744-763.

Winkler, R.L. (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, *67*(337) 187-191.

Yao, Q., & Tong, H. (1996). Asymmetric least squares regression estimation: a nonparametric approach. *Journal of Nonparametric Statistics*, *6*(2-3) 273-292.

Ziegel, J.F. (2016). Coherence and elicitability. *Mathematical Finance*, *26*(4) 901-918.