The Annals of Applied Statistics
2017, Vol. 11, No. 4, 1833–1874
https://doi.org/10.1214/17-AOAS1041

ELICITABILITY AND BACKTESTING: PERSPECTIVES FOR BANKING REGULATION[1]

BY NATALIA NOLDE[2] AND JOHANNA F. ZIEGEL[3]

University of British Columbia and University of Bern

Conditional forecasts of risk measures play an important role in internal risk management of financial institutions as well as in regulatory capital calculations. In order to assess forecasting performance of a risk measurement procedure, risk measure forecasts are compared to the realized financial losses over a period of time and a statistical test of correctness of the procedure is conducted. This process is known as backtesting. Such traditional backtests are concerned with assessing some optimality property of a set of risk measure estimates. However, they are not suited to compare different risk estimation procedures. We investigate the proposal of comparative backtests, which are better suited for method comparisons on the basis of forecasting accuracy, but necessitate an elicitable risk measure. We argue that supplementing traditional backtests with comparative backtests will enhance the existing trading book regulatory framework for banks by providing the correct incentive for accuracy of risk measure forecasts. In addition, the comparative backtesting framework could be used by banks internally as well as by researchers to guide selection of forecasting methods. The discussion focuses on three risk measures, Value at Risk, expected shortfall and expectiles, and is supported by a simulation study and data analysis.

1. Introduction. Financial institutions rely on conditional forecasts of risk measures for the purposes of internal risk management as well as regulatory capital calculations. The two ingredients at the heart of risk measurement are the choice of a suitable risk measure and of a forecasting method, with the forecasting method being typically preceded by the choice of a model and estimation method for the (conditional) loss distribution of the underlying portfolio of risky assets. Traditionally, the choice of a risk measure was based on theoretical considerations linked to practical implications. Emmer, Kratz and Tasche (2015) give a recent account of the pros and cons of popular risk measures with an attempt to determine the best risk measure in practice. On the other hand, Cont, Deguest and Scandolo (2010) highlight the need to consider the entire "risk measurement procedure," which includes not just the choice of a risk measure but also how it is then estimated from

Received May 2016; revised March 2017.

[1]Discussed in 10.1214/17-AOAS1041A, 10.1214/17-AOAS1041B, 10.1214/17-AOAS1041C, 10.1214/17-AOAS1041D, 10.1214/17-AOAS1041E; rejoinder at 10.1214/17-AOAS1041F.

[2]Supported by the Natural Sciences and Engineering Research Council of Canada.

[3]Supported by the Swiss National Foundation Grant 152609.

Key words and phrases. Forecasting, backtesting, elicitability, risk measurement procedure, Value at Risk, expected shortfall, expectiles.

the data. In particular, the notion of robustness as sensitivity to outliers is used to compare several risk measurement procedures. In the risk management context, this should also be balanced with robustness to deviations from model assumptions as well as responsiveness or sensitivity to tail events. Davis (2016) introduces a notion of consistency of risk measures and discusses how this is relevant in the context of financial risk management.

The performance of a (trading book) risk measurement procedure can be monitored over time via a comparison of realized losses with risk measure forecasts, a process known as backtesting; see, for example, Christoffersen (2003) and McNeil, Frey and Embrechts (2005). Based on results of a backtest, the risk measurement procedure is deemed as adequate or not. Traditional backtests perform a statistical test for the null hypothesis:

$$H_0: \quad \text{“The risk measurement procedure is correct.”}$$

If the null hypothesis is not rejected, the risk measurement procedure is considered as adequate. For Value at Risk (VaR), the Bank for International Settlements [(2013), pages 103–108] has devised a three-zone approach based on a binomial test for the number of exceedances over the VaR threshold. Traditional backtests are concerned with assessing an optimality property of a set of risk measure estimates; for details, see Section 2.2. They are not suited to *compare* different risk estimation procedures, and they may be insensitive with respect to increasing information sets; examples of this fact are provided in Holzmann and Eulert (2014), Davis (2016). Moreover, traditional backtests may not provide banks with the right incentive of developing procedures which aim for accuracy of risk measure forecasts; for an illustration, see Section A of the online supplement [Nolde and Ziegel (2017)] (abbreviated “OS” in the sequel). In this simulation-based example, we show how optimization with respect to the test statistic of a traditional backtest may lead to unreasonable ordering of forecasting procedures.

In view of the anticipated revised standardized approach, which “should provide a credible fallback in the event that a bank’s internal market risk model is deemed inadequate” [Bank for International Settlements (2013), pages 5–6], Fissler, Ziegel and Gneiting (2016) have recently proposed to replace traditional backtests by comparative backtests based on strictly consistent scoring functions. Comparative backtests also naturally lead to a three-zone approach, which will be described in detail in Section 2.3. Furthermore, they allow for conservative tests and are sensitive with respect to increasing information sets. Roughly, this means that a risk measurement procedure that correctly incorporates more risk factors will always be preferred over a simpler procedure that uses less information. However, comparative backtests necessitate an *elicitable* risk measure. Examples of elicitable risk measures are VaR and expectiles, while expected shortfall (ES) is not elicitable. However, ES turns out to be jointly elicitable with VaR, which allows for comparative backtests also for ES; for details and a literature review on elicitable risk measures, see Section 2.1.

The paper raises the point of distinguishing between traditional backtesting (current regulatory practice) and comparative backtesting. We highlight the deficiency of the former in giving financial institutions the right incentive for forecast accuracy, and argue that the existing regulatory framework can be enhanced by inclusion of comparative backtesting. On the methodological side, we show that traditional backtesting can be formalized in the form of conditional calibration tests, which provide a unifying framework for many of the existing backtests of popular risk measures. This contributes to our understanding of those often ad hoc procedures and allows us to view them as part of a bigger picture. The paper then provides a detailed investigation of the proposal of comparative backtests.

In our discussion of traditional and comparative backtests, we are focusing on the following three risk measures: VaR, a popular risk measure that is elicitable; expectiles, the only coherent and elicitable risk measures; and ES, a coherent and comonotonically additive risk measure, which is jointly elicitable together with VaR, and which is the new standard measure in banking regulation. VaR at level $\alpha \in (0, 1)$, denoted VaR_α, of a random variable X is defined as

$$\mathrm{VaR}_\alpha(X) = \inf\bigl\{x \mid F_X(x) \geq \alpha\bigr\},$$

where F_X is the cumulative distribution function of X. From the statistical perspective, VaR_α is simply the α-quantile of the underlying distribution, assuming the quantile is single-valued. Positive values of X are interpreted as losses in this manuscript; hence we are interested in VaR_α for values of α close to one. The Bank for International Settlements [(2013), pages 103–108] specifically requests VaR_α values for $\alpha = 0.99$, which we refer to as the standard Basel VaR level. ES of an integrable random variable X at level $\nu \in (0, 1)$ is given by

$$\mathrm{ES}_\nu(X) = \frac{1}{1-\nu}\int_\nu^1 \mathrm{VaR}_\alpha(X)\,d\alpha.$$

The Bank for International Settlements (2014) proposes $\nu = 0.975$ as the standard Basel ES level, as $\mathrm{ES}_{0.975}$ should yield a similar magnitude of risk as $\mathrm{VaR}_{0.99}$ under the standard normal distribution. As introduced by Newey and Powell (1987), the τ-expectile $e_\tau(X)$ of X with finite mean is the unique solution $x = e_\tau(X)$ to the equation

$$\tau\int_x^\infty (y-x)\,\mathrm{d}F_X(y) = (1-\tau)\int_{-\infty}^x (x-y)\,\mathrm{d}F_X(y). \tag{1.1}$$

As shown in Bellini et al. (2014), Ziegel (2016), τ-expectiles are elicitable coherent risk measures for $\tau \in [1/2, 1)$. Expectiles generalize the expectation just as quantiles generalize the median. Considering the level, $\tau = 0.99855$ leads to a comparable magnitude of risk as $\mathrm{VaR}_{0.99}$ and $\mathrm{ES}_{0.975}$ under the standard normal distribution; see Bellini and Di Bernardino (2017).

The paper is organized as follows. Section 2 contains a theoretical discussion of backtesting risk measures. In Section 2.1 we define the notion of elicitability,

introduce identifiability and review characterizations of consistent scoring functions for VaR, expectiles and (VaR, ES). In Section 2.2 we define what we mean by a calibrated risk measurement procedure and describe how this concept is related to the notion of calibration of Davis (2016) and to traditional backtests in general. We move on to comparative backtests in Section 2.3, where we also explain the comparative three-zone approach. Section 2.4 discusses the choice of the scoring function. Section 3 contains numerical studies of the proposed backtesting methodologies. We first review some of the existing approaches to forecasting risk measures in Section 3.1. A simulation study is described in Section 3.2, while an application to the returns on the NASDAQ Composite index is presented in Section 3.3. Section 4 concludes the paper with a summary and a discussion of the findings, in particular, in relation to banking regulation. Section B in the OS contains the necessary background material for computing and estimation of expectiles, and gives a derivation of an extreme value-based estimator; some of the results here are of interest in their own right. Technical results on the characterization of consistent scoring functions with positive-homogeneous score differences are delegated to Section C of the OS. Finally, Section D of the OS reports results of a simulation study which investigates the performance of backtesting procedures in the setting where the out-of-sample size is small.

2. Backtesting of risk measures.

2.1. *Preliminaries.* A risk measure ρ is usually defined on some space of random variables. In this paper, we only consider risk measures that are law-invariant; that is, two random variables X and Y with the same distribution $\mathcal{L}(X) = \mathcal{L}(Y)$ are assigned the same value of ρ. Therefore, we view a risk measure ρ as a map from some collection of probability distributions $\mathcal{P}$ to the real line $\mathbb{R}$. Then the risk of X with distribution $\mathcal{L}(X)$ is $\rho(\mathcal{L}(X))$. In some instances, where no confusion can arise, we abuse notation and write $\rho(X)$ instead of $\rho(\mathcal{L}(X))$. Let $\Theta = (\rho_1, \dots, \rho_k)$ be a vector of $k \geq 1$ risk measures.

DEFINITION 1. A scoring function $S : \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}$ is called *strictly consistent* for Θ with respect to $\mathcal{P}$ if

$$\mathbb{E}\big(S\big(\Theta\big(\mathcal{L}(X)\big), X\big)\big) < \mathbb{E}\big(S(r, X)\big) \tag{2.1}$$

for all $r = (r_1, \dots, r_k) \neq \Theta(\mathcal{L}(X)) = (\rho_1(\mathcal{L}(X)), \dots, \rho_k(\mathcal{L}(X)))$ and all X with distribution $\mathcal{L}(X) \in \mathcal{P}$. The scoring function S is *consistent* if equality is allowed in (2.1). The vector of risk measures Θ is called *elicitable* with respect to $\mathcal{P}$ if there exists a strictly consistent scoring function for it.

Elicitability is useful for model selection, estimation, generalized regression, forecast ranking, and, as we will detail in this paper, allows for comparative backtesting. Elicitable functionals were already studied in the thesis of Osband (1985),

although the terminology was coined by Lambert, Pennock and Shoham (2008). A comprehensive literature review on elicitability can be found in Gneiting (2011), where particular emphasis is on the case $k = 1$. Recent advances on the case $k \geq 2$ can be found in Frongillo and Kash (2015), Fissler and Ziegel (2016).

The question of elicitability of risk measures has recently received considerable attention. All available results in the case $k = 1$ are based on the simple but powerful observation that a necessary requirement of elicitability is convex level sets in a distributional sense [Osband (1985)]; see also Gneiting (2011), Theorem 6. Weber (2006) was the first to study risk measures with convex level sets. Bellini and Bignozzi (2015) used his results to study elicitability for the broad class of monetary risk measures. Under weak regularity assumptions, they show that elicitable monetary risk measures are so-called shortfall risk measures [Föllmer and Schied (2002)]. For more specific classes of risk measures, such as coherent, convex or distortion risk measures, the same result can be shown without any additional regularity assumptions [Ziegel (2016), Delbaen et al. (2016), Kou and Peng (2016), Wang and Ziegel (2015)]. While expected shortfall is itself not elicitable, Fissler and Ziegel (2016) have shown that the pair $\Theta = (\mathrm{VaR}_\alpha, \mathrm{ES}_\alpha)$ is elicitable; see also Acerbi and Szekely (2014).

The classes of (strictly) consistent scoring functions for VaR_α, τ-expectiles and $(\mathrm{VaR}_\upsilon, \mathrm{ES}_\upsilon)$ have been characterized. The following three propositions state sufficient conditions for (strict) consistency. Under mild regularity assumptions given in the cited literature and up to equivalence, these conditions are also necessary. Here, two scoring functions are called *equivalent* if their difference is a function of the realization $x \in \mathbb{R}$ only. Let $\mathcal{P}_0$ denote the class of all Borel-probability distributions on $\mathbb{R}$, and let $\mathcal{P}_1 \subseteq \mathcal{P}_0$ denote the class of all distributions with finite mean.

PROPOSITION 1 [Thomson (1979), Saerens (2000)]. *All scoring functions of the form*

$$S(r, x) = \big(1 - \alpha - \mathbb{1}\{x > r\}\big)G(r) + \mathbb{1}\{x > r\}G(x), \tag{2.2}$$

where G is an increasing function on $\mathbb{R}$, are consistent for VaR_α, $\alpha \in (0, 1)$, with respect to $\mathcal{P}_0$. The scoring functions of the above form are strictly consistent for VaR_α with respect to $\mathcal{P}' \subseteq \mathcal{P}_0$ if G is strictly increasing, $G(X)$ is integrable for all X with distribution in $\mathcal{P}'$, and all distributions in $\mathcal{P}'$ have a unique α-quantile.

PROPOSITION 2 [Gneiting (2011)]. *All scoring functions of the form*

$$\begin{aligned} S(r, x) = {} & \mathbb{1}\{x > r\}(1 - 2\tau)\big(\phi(r) - \phi(x) - \phi'(r)(r - x)\big) \\ & - (1 - \tau)\big(\phi(r) - \phi'(r)(r - x)\big), \end{aligned} \tag{2.3}$$

where ϕ is a convex function with subgradient ϕ', are consistent for the τ-expectile, $\tau \in (0, 1)$, with respect to $\mathcal{P}_1$. If ϕ is strictly convex, then the scoring functions of the above form are strictly consistent for the τ-expectile relative to the class $\mathcal{P}' \subseteq \mathcal{P}_1$ such that $\phi(X)$ is integrable for all X with distribution in $\mathcal{P}'$.

PROPOSITION 3 [Fissler and Ziegel (2016)]. *All scoring functions of the form*

$$
\begin{aligned}
S(r_1, r_2, x) = {} & \mathbb{1}\{x > r_1\}\big(-G_1(r_1) + G_1(x) - G_2(r_2)(r_1 - x)\big) \\
& + (1 - \nu)\big(G_1(r_1) - G_2(r_2)(r_2 - r_1) + \mathcal{G}_2(r_2)\big),
\end{aligned} \tag{2.4}
$$

where G_1 is an increasing function, $\mathcal{G}_2' = G_2$ and $\mathcal{G}_2$ is increasing and concave, are consistent for $(\mathrm{VaR}_\nu, \mathrm{ES}_\nu)$, $\nu \in (0, 1)$, *with respect to $\mathcal{P}_1$. If $\mathcal{G}_2$ is strictly increasing and strictly concave, then the above scoring functions are strictly consistent with respect to the subclass $\mathcal{P}' \subseteq \mathcal{P}_1$ of the distributions $P \in \mathcal{P}_1$ with a unique ν-quantile and such that $G_1(X)$ is integrable when X has distribution P.*

In risk management applications, it may be useful to allow only for strictly positive risk measure predictions. As shown in Section 2.4, this opens up the possibility for attractive choices of homogeneous scoring functions in the above propositions. If $r \in (0, \infty)$ is assumed in (2.2) or (2.3), then, for strict consistency, we only need that G or ϕ are defined on $(0, \infty)$, and that they are strictly increasing or strictly convex on this domain, respectively. In the case of (2.2), this can be checked by a fairly straightforward computation. For the claim concerning (2.3), it is useful to consider the decomposition of the score difference derived in the proof of Gneiting (2011), Theorem 10. Furthermore, it is sufficient to require integrability of $G(X)\mathbb{1}\{X > 0\}$ or $\phi(X)\mathbb{1}\{X > 0\}$ for all X with distribution in $\mathcal{P}'$. If we restrict to predictions with $(r_1, r_2) \in \mathbb{R} \times (0, \infty)$ in (2.4), $\mathcal{G}_2$ only has to be defined on $(0, \infty)$ and has to be strictly increasing and strictly concave on this domain.

Closely connected to elicitability is the concept of identifiability. In fact, for $k = 1$, identifiability implies elicitability under some additional assumptions; see Steinwart et al. (2014). For $k \geq 2$, it is currently unclear whether such a general result holds; see Fissler and Ziegel (2016).

DEFINITION 2. The vector of risk measures Θ is called *identifiable* with respect to $\mathcal{P}$ if there is a function $V : \mathbb{R}^k \times \mathbb{R} \to \mathbb{R}^k$ such that

$$
\mathbb{E}\big(V(r, X)\big) = 0 \quad \Leftrightarrow \quad r = \Theta\big(\mathcal{L}(X)\big)
$$

for all X with distribution $\mathcal{L}(X) \in \mathcal{P}$.

Identification functions are not uniquely defined. In fact, one can multiply any identification function for a functional by a function depending only on the prediction r and taking values in the space of invertible $k \times k$-matrices to obtain another identification function for the same functional.

VaR_α for $\alpha \in (0, 1)$ is identifiable with respect to the class $\mathcal{P}_V \subset \mathcal{P}_0$ of distributions with unique quantiles with identification function

$$
V(r, x) = 1 - \alpha - \mathbb{1}\{x > r\}, \tag{2.5}
$$

the τ-expectile for $\tau \in (0,1)$ is identifiable with respect to $\mathcal{P}_1$ using the identification function

$$V(r,x) = \big|1-\tau-\mathbb{1}\{x>r\}\big|(r-x), \tag{2.6}$$

and $(\mathrm{VaR}_\nu, \mathrm{ES}_\nu)$ for the level $\nu \in (0,1)$ has identification function

$$V(r_1,r_2,x) = \begin{pmatrix} 1-\nu-\mathbb{1}\{x>r_1\} \\ r_1 - r_2 - \dfrac{1}{1-\nu}\mathbb{1}\{x>r_1\}(r_1-x) \end{pmatrix} \tag{2.7}$$

with respect to $\mathcal{P}_1 \cap \mathcal{P}_V$.

2.2. *Calibration and traditional backtests.* We fix the following notation. Suppose that $\Theta = (\rho_1, \dots, \rho_k)$ is an identifiable functional with identification function V with respect to $\mathcal{P}$. Let $\{X_t\}_{t\in\mathbb{N}}$ be a series of negated log-returns adapted to the filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}}$ and $\{R_t\}_{t\in\mathbb{N}}$ a sequence of predictions of Θ, which are $\mathcal{F}_{t-1}$-measurable. Hence the predictions are based on the information about $\{X_t\}_{t\in\mathbb{N}}$ available at time $t-1$ represented by the sigma-algebra $\mathcal{F}_{t-1}$. Let $\mathcal{L}(X_t|\mathcal{F}_{t-1})$ denote the conditional law of X_t given the information $\mathcal{F}_{t-1}$. We assume that all conditional distributions $\mathcal{L}(X_t|\mathcal{F}_{t-1})$ and all unconditional distributions $\mathcal{L}(X_t)$ belong to $\mathcal{P}$ almost surely.

Inspired by the insightful paper of Davis (2016), we give the following definition.

DEFINITION 3. The sequence of predictions $\{R_t\}_{t\in\mathbb{N}}$ is *calibrated for* Θ *on average* if

$$\mathbb{E}\big(V(R_t,X_t)\big) = 0 \qquad \text{for all } t \in \mathbb{N};$$

it is *super-calibrated for* Θ *on average* if $\mathbb{E}(V(R_t,X_t)) \geq 0$ component-wise for all $t \in \mathbb{N}$. The sequence of predictions $\{R_t\}_{t\in\mathbb{N}}$ is *conditionally calibrated for* Θ if

$$\mathbb{E}\big(V(R_t,X_t)|\mathcal{F}_{t-1}\big) = 0 \qquad \text{almost surely, for all } t \in \mathbb{N};$$

it is *conditionally super-calibrated for* Θ if $\mathbb{E}(V(R_t,X_t)|\mathcal{F}_{t-1}) \geq 0$ component-wise, almost surely, for all $t \in \mathbb{N}$. *Sub-calibration* is defined analogously.

If one knows the conditional distributions $\mathcal{L}(X_t|\mathcal{F}_{t-1})$ and strives for the best possible prediction of Θ based on the information in $\mathcal{F}_{t-1}$, it is natural to use

$$\Theta\big(\mathcal{L}(X_t|\mathcal{F}_{t-1})\big) \tag{2.8}$$

as a predictor, which we term the *optimal $\mathcal{F}$-conditional forecast* for Θ. For the same reason, we call $\Theta(\mathcal{L}(X_t))$ the *optimal unconditional forecast*. Recall that we freely abuse notation in using Θ either as a functional defined on a space of random variables or on a space of probability distributions.

Calibration characterizes optimal forecasts in the following sense. The optimal unconditional forecast is the only deterministic forecast that is calibrated for Θ on average. However, there may be other forecasts that are calibrated for Θ on average which are not deterministic and thus different from the optimal unconditional forecast. Likewise, the optimal conditional forecast is the only $\mathcal{F}$-predictable conditionally calibrated forecast for Θ up to almost sure equivalence. It is clear that conditional calibration implies calibration on average by the tower property of conditional expectations, but the converse is generally false. The notions of calibration introduced here are analogous to the notions of cross-calibration for probabilistic forecasts introduced in Strähl and Ziegel (2017).

We have introduced the notions of super- and sub-calibration as they can often be related to over- or under-estimation of the risk measure at hand. However, this depends on the specific identification function, and so some care must be taken. We give details for a correct interpretation for VaR, expectiles and (VaR, ES) in Section 2.2.2.

For simplicity, we focus on one-step ahead predictions in this paper. Clearly, multi-step ahead predictions are equally important. In some instances the same theory and concepts can be transferred from the former case to the latter.

Following Fissler, Ziegel and Gneiting (2016), we call any backtest that considers a null hypothesis of the type "The risk measurement procedure is correct" a *traditional backtest*. Traditional backtests are similar to goodness-of-fit tests, that is, they allow to demonstrate that the risk measurement procedure under consideration is making incorrect predictions, if the respective null hypothesis can be rejected. Despite the somewhat misleading terminology that a traditional backtest is *passed* if the null hypothesis is not rejected, this does *not* mean that, in this case, one can be sure that the null hypothesis is correct (with a prespecified small probability of error), as this would necessitate that we control the power of the test explicitly. This can virtually never be done, as the alternative is too broad; see also Bank for International Settlements (2013), pages 103–105. As argued by Fissler, Ziegel and Gneiting (2016), these issues may put the use of the traditional backtest in regulatory frameworks in question. However, they may be useful for model verification just as goodness-of-fit tests have their established role in statistics.

Testing the null hypothesis

$$H_0: \quad \text{The sequence of predictions } \{R_t\}_{t\in\mathbb{N}} \text{ is calibrated for } \Theta \text{ on average} \tag{2.9}$$

amounts to performing a traditional backtest. We describe here how tests for average calibration can be constructed, but we do not implement them because the stronger notion of conditional calibration appears more adequate in a dynamic risk management context. In our data example in Section 3.3, for the more flexible models, the null hypothesis of conditional calibration cannot be rejected, which indicates that testing for average calibration is superfluous. However, there may be situations where achieving average calibration is already difficult, and then the following tests may be useful.

Given a series of observations $\{X_t\}_{t=1,\dots,n}$ and forecasts $\{R_t\}_{t=1,\dots,n}$, we define $\overline{V}_n := (1/n)\sum_{t=1}^n V(R_t, X_t)$. Let $\hat{\Sigma}_n$ be a heteroscedasticity and autocorrelation consistent (HAC) estimator of the asymptotic covariance matrix $\Sigma_n = \mathrm{cov}(\sqrt{n}\overline{V}_n)$ (see Andrews, 1991). Then one can hope that $\sqrt{n}\hat{\Sigma}_n^{-1/2}\overline{V}_n$ is asymptotically standard normal under suitable assumptions on the identification function and the data-generating process. For $k = 1$, sufficient mixing assumptions are detailed in Giacomini and White [(2006), Theorem 4], but a multivariate generalization of this result remains to be worked out. Giacomini and White [(2006), Theorem 4] show that, for $k = 1$, the test is consistent against the alternative $|\mathbb{E}(\overline{V}_n)| \geq \delta > 0$ for all n sufficiently large for any $\delta > 0$.

Conditional calibration is a stronger notion than average calibration, and it appears more natural in a dynamic risk management context. A traditional backtest for conditional calibration considers the null hypothesis

$$H_0: \quad \text{The sequence of predictions } \{R_t\}_{t\in\mathbb{N}} \text{ is conditionally calibrated for } \Theta. \tag{2.10}$$

The requirement $\mathbb{E}(V(R_t, X_t)|\mathcal{F}_{t-1}) = 0$, almost surely, is equivalent to stating that $\mathbb{E}(h_t' V(R_t, X_t)) = 0$ for all $\mathcal{F}_{t-1}$-measurable $\mathbb{R}^k$-valued functions h_t. Following Giacomini and White (2006), we consider an $\mathcal{F}$-predictable sequence $\{\mathbf{h}_t\}_{t\in\mathbb{N}}$ of $q \times k$-matrices $\mathbf{h}_t$ called *test functions* to construct a Wald-type test statistic:

$$T_1 = n\left(\frac{1}{n}\sum_{t=1}^n \mathbf{h}_t V(R_t, X_t)\right)' \widehat{\Omega}_n^{-1}\left(\frac{1}{n}\sum_{t=1}^n \mathbf{h}_t V(R_t, X_t)\right), \tag{2.11}$$

where

$$\widehat{\Omega}_n = \frac{1}{n}\sum_{t=1}^n \big(\mathbf{h}_t V(R_t, X_t)\big)\big(\mathbf{h}_t V(R_t, X_t)\big)'$$

is a consistent estimator of the variance of the q-vector $\mathbf{h}_t V(R_t, X_t)$. Ideally, the parameter q should be chosen such that the rows of $\mathbf{h}_t$ generate $\mathcal{F}_{t-1}$. In applications, the choice of the test functions is motivated by the principle that they should represent the most important information available at time point $t - 1$. In our simulation study, we obtained good results with $q = 1$ or $q = 2$; for further details, see Section 3.2.2. We call this type of traditional backtests *conditional calibration tests*. In cases where $\mathbf{h}_t = 1$, we refer to these tests as *simple* conditional calibration tests. Theorem 1 in Giacomini and White (2006) states that, under the null hypothesis (2.10), $T_1 \xrightarrow{\mathrm{d}} \chi_q^2$ as $n \to \infty$, subject to certain assumptions on the data-generating process $\{X_t\}_{t\in\mathbb{N}}$ and test function sequence $\{\mathbf{h}_t\}_{t\in\mathbb{N}}$. This asymptotic result justifies a level η test which rejects H_0 when $T_1 > \chi^2_{q,1-\eta}$, where $\chi^2_{q,1-\eta}$ denotes the $1 - \eta$ quantile of the χ_q^2 distribution. Giacomini and White [(2006), Theorem 3] provide conditions such that $T_1 \xrightarrow{\mathrm{d}} \chi_q^2$ as $n \to \infty$ for multi-step ahead

predictions, while Theorem 2 of Giacomini and White (2006) considers consistency of the test against global alternatives. The theorems of Giacomini and White (2006) are formulated in terms of score differences and not identification functions, but their proofs solely rely on the martingale difference property of $\mathbf{h}_t V(R_t, X_t)$ and can thus be applied in our context.

Commonly used backtests for VaR_α and ES_ν are closely related to conditional calibration tests for specific choices of the test functions $\mathbf{h}_t$. In fact, choosing $\mathbf{h}_t = 1$ in the case of VaR_α, the conditional calibration test for VaR_α is closely related to the standard backtest for VaR_α based on the number of VaR exceedances [Bank for International Settlements (2013), pages 103–108]. In the case of ES_ν, the conditional calibration test for $(\mathrm{VaR}_\nu, \mathrm{ES}_\nu)$ is related to the backtest for ES_ν of McNeil and Frey (2000) based on exceedance residuals. We give further details in Examples 1, 2 and 3 below.

The notion of a calibrated risk measure (or statistic) of Davis (2016) is closely related to our notion of a calibrated sequence of predictions. Davis (2016) considers which risk measures are calibrated for which classes of models; that is, he attempts to characterize the largest class of data-generating processes such that $\overline{V}_n$ goes to zero a.s. as $n \to \infty$ if $\{R_t\}_{t\in\mathbb{N}}$ is a sequence of optimal conditional forecasts for the risk measure. It turns out that for quantiles only minimal assumptions are necessary, whereas assumptions need to be stronger to work with the mean, for example. The focus of our work is more statistical. Choosing $\mathcal{F}$-predictable test functions $\mathbf{h}_t$ encoding the available information at time point $t-1$, we investigate whether and how it is possible to test in finite samples that the sequence $\{R_t\}_{t\in\mathbb{N}}$ is conditionally calibrated.

2.2.1. *One-sided calibration tests.* In certain situations, it may be meaningful to assess super- or sub-calibration. For example, the standard backtest for VaR_α described in the Bank for International Settlements [(2013), pages 103–108], is a test for conditional super-calibration. This is due to the fact that over-estimation of VaR_α is not a problem as far as the regulator is concerned. Holding more capital than minimally required should always be allowed.

Suppose we wish to test the hypothesis of conditional super-calibration that $\mathbb{E}[V(R_t, X_t)|\mathcal{F}_{t-1}] \geq \mathbf{0}$ component-wise for all t; that is, in the case of a k-variate risk measure, we are interested in $H_0 = \bigcap_{i=1}^k H_{0,i}$, where

$$\mathrm{H}_{0,i}: \quad \mathbb{E}\big[V_i(R_t, X_t)|\mathcal{F}_{t-1}\big] \geq 0 \qquad \text{for all } t, i = 1, \ldots, k.$$

For each component i of the risk measure, let $\mathbf{h}_{i,t} = (h_{i,t,1}, \ldots, h_{i,t,q_i})$ be an $\mathcal{F}_{t-1}$-measurable $(q_i \times 1)$-vector of non-negative test functions. If $h_{i,t,1}, \ldots, h_{i,t,q_i}$ generate $\mathcal{F}_{t-1}$, then $H_{0,i} = \bigcap_{\ell=1}^{q_i} H_{0,i,\ell}$, where

$$\mathrm{H}_{0,i,\ell}: \quad \mathbb{E}\big[V_i(R_t, X_t)h_{i,t,\ell}\big] \geq 0 \qquad \text{for all } t, i = 1, \ldots, k, \ell = 1, \ldots, q_i.$$

We combine all of the test functions into a $(q \times k)$ matrix $\mathbf{h}_t$ with $q = \sum_{i=1}^{k} q_i$, which has the following structure:

$$\mathbf{h}_t = \begin{pmatrix} \mathbf{h}_{1,t} & 0 & \cdots & 0 \\ 0 & \mathbf{h}_{2,t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{h}_{k,t} \end{pmatrix}.$$

Setting $Z_t = \mathbf{h}_t V(R_t, X_t)$, the above hypothesis of conditional super-calibration can alternatively be expressed as $H_0 = \bigcap_{m=1}^{q} H_{0,m}$ with $H_{0,m} : \mathbb{E}(Z_{t,m}) \geq 0$ for all $t, m = 1, \ldots, q$.

From the proof of Giacomini and White [(2006), Theorems 1 and 3] it follows that, under H_0 given at (2.10),

$$T_2 = (T_{2,1}, \ldots, T_{2,q})' = \sqrt{n}^{-1} \widehat{\Omega}_n^{-1/2} \sum_{t=1}^{n} Z_t \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0}, I_q), \qquad n \to \infty, \tag{2.12}$$

where I_q denotes the $(q \times q)$ identity matrix. Hence we can obtain an asymptotic test for $H_{0,m}$ with the p-value given by $\pi_m = \Phi(\sqrt{n}^{-1} (\widehat{\Omega}_n)_{mm}^{-1/2} \sum_{t=1}^{n} Z_{t,m})$, $m = 1, \ldots, q$; that is, π_m is the (asymptotic) probability of obtaining a more extreme outcome than the one observed, assuming the null hypothesis $H_{0,m}$ is true. Let $\pi_{(1)}, \ldots, \pi_{(q)}$ be the ordered p-values. The classical Bonferroni multiple test procedure rejects the global null hypothesis H_0 if the smallest of the p-values $\pi_{(1)} < \eta/q$, where η is the desired level of the (global) test. As an alternative, following Hommel (1983), we obtain a level η test by rejecting the global hypothesis H_0 if for at least one m we have

$$\pi_{(m)} \leq \frac{m\eta}{q C_q}, \qquad C_q = \sum_{r=1}^{q} 1/r, m = 1, \ldots, q. \tag{2.13}$$

Hommel's rejection rule has the advantage of allowing to detect situations with both small effects in many components and with large effects in few components. Other testing procedures in this context could also be used.

2.2.2. *Examples.*

EXAMPLE 1. Christoffersen (1998) calls a sequence of VaR_α forecasts efficient with respect to $\mathcal{F}$ if

$$\mathbb{E}\big[\mathbb{1}\{X_t > R_t\} | \mathcal{F}_{t-1}\big] = 1 - \alpha \qquad \text{almost surely, } t = 1, 2, \ldots.$$

This requirement is the same as the one of conditional calibration of $\{R_t\}_{t \in \mathbb{N}}$ by (2.5). In fact, the dynamic quantile test of Kuester, Mittnik and Paolella (2006) [see also Christoffersen (1998), Engle and Manganelli (2004)] has similarities to

a conditional calibration test. In analogy to their test, it is natural to consider test functions

$$\mathbf{h}_t = \big(1, V(r_{t-1}, x_{t-1}), \ldots, V(r_{t-p}, x_{t-p}), r_t\big)'$$

for $p \geq 1$. This is also in line with the suggestion in Giacomini and White (2006) who use $\mathbf{h}_t = (1, V(r_{t-1}, x_{t-1}))'$.

The standard backtest for VaR_α specified in the Basel documents [Bank for International Settlements (2013), pages 103–108] uses the test statistic

$$\beta = \sum_{t=1}^{n} \mathbb{1}\{X_t > R_t\},$$

which is the number of exceedances over the estimated VaR_α, denoted R_t, for time point t. Under the null hypothesis (2.10) of conditionally calibrated VaR_α-forecasts, for one-step ahead forecasts, β is a binomial random variable with parameters n and $1-\alpha$; see Rosenblatt (1952), Diebold, Gunther and Tay (1998), Davis (2016). It is remarkable that this result holds under essentially no assumptions on $\{X_t\}_{t\in\mathbb{N}}$ or $\{R_t\}_{t\in\mathbb{N}}$. However, when moving away from one-step ahead forecasts to multi-step ahead forecasts, things become more intricate and one has to resort to general limit theorems such as presented above for testing if β has mean $n(1-\alpha)$. This test is a test for conditional super-calibration with $\mathbf{h}_t = 1$ because, for VaR_α, we obtain using (2.5)

$$\begin{aligned} T_3 &:= \sum_{t=1}^{n} \mathbf{h}_t V(R_t, X_t) = \sum_{t=1}^{n} \big(\mathbb{1}\{X_t \leq R_t\} - \alpha\big) \\ &= \sum_{t=1}^{n} \big(\mathbb{1}\{X_t > R_t\} - (1-\alpha)\big) = -\big(\beta - n(1-\alpha)\big), \end{aligned}$$

and thus testing the null hypothesis that β has mean less or equal to $n(1-\alpha)$ is equivalent to testing that T_3 has mean greater or equal to zero. This null hypothesis says that the conditional VaR predictions are at least as large as the true conditional VaR. Assuming that it is an incentive of a bank to state VaR estimates that tend to be lower than the true ones, a more prudent null hypothesis from the viewpoint of a regulator would be the opposite one-sided hypothesis that the conditional VaR predictions are at most as large as the true conditional VaR, that is, a test for conditional sub-calibration.

For one-step ahead predictions, alternatively to theory presented in this section, one can exploit the fact that the exceedance indicators $\mathbb{1}\{X_t > R_t\}$, $t = 1, \ldots, n$ at the boundary of the null hypothesis, are independent Bernoulli random variables with success probability $1-\alpha$, which allows for an exact test rather than an asymptotic one.

EXAMPLE 2. We consider the vector of risk measures $\Theta(\mathcal{L}(X)) = (\rho_1(\mathcal{L}(X)), \rho_2(\mathcal{L}(X))) = (\mathrm{VaR}_\nu(X), \mathrm{ES}_\nu(X))$ for some $\nu \in (0,1)$. Let $r_{1,t}$ and $r_{2,t}$ denote forecasts of $\mathrm{VaR}_\nu(X_t)$ and $\mathrm{ES}_\nu(X_t)$, respectively. Assuming $X_t = \mu_t + \sigma_t Z_t$, where μ_t and σ_t are $\mathcal{F}_{t-1}$-measurable and the Z_t's form an independent and identically distributed (i.i.d.) sequence of random variables with zero mean and variance one, for backtesting ES, McNeil and Frey (2000) introduced the following test statistic based on exceedance residuals:

$$T_4 = \frac{1}{\#\{t : X_t > r_{1,t}\}} \sum_{t=1}^{n} \frac{X_t - r_{2,t}}{\sigma_t} \mathbb{1}\{X_t > r_{1,t}\}. \tag{2.14}$$

It turns out that the ES backtest of McNeil and Frey (2000) is closely related to a conditional calibration test as follows. For n reasonably large, we have that $\#\{t : x_t > r_{1,t}\}/n \approx 1 - \nu$. Therefore, for the test statistic T_4 in (2.14), we obtain

$$T_4 \approx \frac{1}{n} \sum_{t=1}^{n} \frac{1}{1-\nu} \frac{x_t - r_{2,t}}{\sigma_t} \mathbb{1}\{x_t > r_{1,t}\} = \frac{1}{n} \sum_{t=1}^{n} \mathbf{h}_t V(r_{1,t}, r_{2,t}, x_t)$$

with $\mathbf{h}_t = \sigma_t^{-1}((r_{2,t} - r_{1,t})/(1-\nu), 1)$. Replacing σ_t by an estimate $\hat{\sigma}_t$ is natural when considering the test of McNeil and Frey (2000) as a conditional calibration test. The estimated volatility $\hat{\sigma}_t$ is then simply a part of the $\mathcal{F}_{t-1}$-measurable test function sequence $\{\mathbf{h}_t\}_{t \in \mathbb{N}}$ that supposedly encodes the relevant information of $\mathcal{F}_{t-1}$. Of course, this test is only reasonable if σ_t is estimated as part of the forecasting model with the information at time point $t-1$. The recently proposed backtests for ES of Acerbi and Szekely (2014) are in the same spirit as the test of McNeil and Frey (2000).

The backtest for ES suggested by Costanzino and Curran (2015) tests if the whole tail of the distribution beyond the VaR_ν-level has been estimated correctly. Strictly speaking, the test is therefore not a test for the accuracy of a sequence of point forecasts for $(\mathrm{VaR}_\nu, \mathrm{ES}_\nu)$ but rather a test for the accuracy of a sequence of probabilistic forecasts for tomorrow's loss distribution with emphasis on the left tail. Other tests in this spirit but of comparative type can be found in Gneiting and Ranjan (2011).

As ES is only identifiable jointly with VaR, one has to be careful when formulating a one-sided test for ES. Let $(r_1^*, r_2^*) = \Theta(\mathcal{L}(X))$. Then it holds for all (r_1, r_2) that

$$r_2^* - r_2 \le \mathbb{E}V_2(r_1, r_2, X) \le r_2^* - r_2 + \frac{\nu - F(r_1)}{1-\nu}(r_1^* - r_1).$$

This shows that, similarly to the VaR case, testing the null hypothesis of subcalibration for the ES component $\mathbb{E}V_2(r_1, r_2, X) \le 0$ is equivalent to testing that $r_2^* \le r_2$. Hence the test of conditional sub-calibration of (VaR, ES) is a test that the conditional VaR and ES predictions are at least as large as their optimal conditional predictions. The Hommel's procedure described in Section 2.2.1 can then be applied with p-value $\pi_m = 1 - \Phi(T_{2,m})$, where the $T_{2,m}$'s are defined in (2.12).

EXAMPLE 3. One could conceive a backtesting framework for expectiles as well in a similar spirit to the ES backtesting procedure proposed by McNeil and Frey (2000). Assuming, as in the example above, that $X_t = \mu_t + \sigma_t Z_t$, where μ_t and σ_t are $\mathcal{F}_{t-1}$-measurable and the Z_t's are i.i.d. with zero mean and variance one, the conditional τ-expectile satisfies

$$e_\tau(X_t \mid \mathcal{F}_{t-1}) = \mu_t + \sigma_t e_\tau(Z_t),$$

and we see that the residuals

$$\frac{X_t - e_\tau(X_t \mid \mathcal{F}_{t-1})}{\sigma_t} = Z_t - e_\tau(Z_t)$$

form an i.i.d. sequence of random variables with zero τ-expectile. This implies that $V(e_\tau(Z_t), Z_t)$ with V given at (2.6) is an i.i.d. sequence of random variables with mean zero, which can be tested using a bootstrap [as in Efron and Tibshirani (1993), Section 16.4]. Here it is necessary to replace the true volatility σ_t by an estimate. This is analogous to the suggestion of McNeil and Frey (2000) for ES. Noticing that the identification function for expectiles at (2.6) is positively 1-homogeneous, we obtain that

$$\mathbb{E}V\big(e_\tau(Z_t), Z_t\big) = \mathbb{E}V\big(e_\tau(X_t), X_t\big)\sigma_t^{-1} = 0.$$

This equality suggests that it is natural to perform a conditional calibration test for expectiles with test function $\mathbf{h}_t = \hat{\sigma}_t^{-1}$ and test statistic T_1 given at (2.11). This yields a valid asymptotic test under the assumptions in Giacomini and White (2006), Theorem 1. These assumptions are weaker than the model assumption $X_t = \mu_t + \sigma_t Z_t$.

In the case of expectiles, as in the case of VaR, a test for conditional supercalibration assesses the null hypothesis that all conditional expectile estimates are at least as large as the true conditional expectile.

2.3. *Elicitability, forecast dominance and comparative backtests.* Suppose now that the functional $\Theta = (\rho_1, \ldots, \rho_k)$ is elicitable with respect to $\mathcal{P}$. Let $\{X_t\}_{t\in\mathbb{N}}$ be a series of negated log-returns adapted to the filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\in\mathbb{N}}$ as well as to the filtration $\mathcal{F}^* = \{\mathcal{F}_t^*\}_{t\in\mathbb{N}}$. Let $\{R_t\}_{t\in\mathbb{N}}$ and $\{R_t^*\}_{t\in\mathbb{N}}$ be two sequences of predictions of Θ, which are $\mathcal{F}$ and $\mathcal{F}^*$-predictable, respectively. We assume that all conditional distributions $\mathcal{L}(X_t|\mathcal{F}_{t-1})$, $\mathcal{L}(X_t|\mathcal{F}_{t-1}^*)$ and all unconditional distributions $\mathcal{L}(X_t)$ belong to $\mathcal{P}$ almost surely. We refer to the predictions $\{R_t^*\}_{t\in\mathbb{N}}$ as the standard procedure, while $\{R_t\}_{t\in\mathbb{N}}$ is the internal model. The two filtrations $\mathcal{F}$ and $\mathcal{F}^*$ acknowledge the fact that the internal model and the standard model may be based on different information sets. For example, one model may include more risk factors than the other, or certain expert opinion may be used to adjust one model but not the other.

DEFINITION 4. Let S be a consistent scoring function for Θ with respect to $\mathcal{P}$. Then $\{R_t\}_{t\in\mathbb{N}}$ *S-dominates* $\{R_t^*\}_{t\in\mathbb{N}}$ *(on average)* if

$$\mathbb{E}\big(S(R_t, X_t) - S\big(R_t^*, X_t\big)\big) \le 0 \qquad \text{for all } t \in \mathbb{N}.$$

Furthermore, $\{R_t\}_{t\in\mathbb{N}}$ *conditionally S-dominates* $\{R_t^*\}_{t\in\mathbb{N}}$ if

$$\mathbb{E}\big(S(R_t, X_t) - S\big(R_t^*, X_t\big)|\mathcal{F}_{t-1}^*\big) \le 0 \qquad \text{almost surely, for all } t \in \mathbb{N}. \tag{2.15}$$

The definition of conditional dominance is asymmetric in terms of the role of the standard procedure and the internal procedure. The standard procedure and the information $\mathcal{F}^*$ it is based on are considered as a benchmark of predictive ability, which is why we condition on $\mathcal{F}_{t-1}^*$ and not on $\mathcal{F}_{t-1}$. Any method that dominates the benchmark has superior predictive ability relative to this benchmark.

Clearly, conditional S-dominance implies S-dominance on average. Ehm et al. [(2016), Definition 2] introduced the notion of dominance of one sequence of predictions over the other if one S-dominates the other on average for all consistent scoring functions S for Θ. The notion of dominance is a strong one; that is, in the data examples of Ehm et al. (2016) it was almost never observed that one forecast dominates the other. This makes the concept difficult to employ in an applied decision-making context. Furthermore, currently, a clear theoretical understanding of the notion of dominance remains elusive.

There are several reasons why the predictions $\{R_t\}_{t\in\mathbb{N}}$ should be preferred over $\{R_t^*\}_{t\in\mathbb{N}}$ if the former dominates the latter. First, comparison of forecasts with respect to the described dominance relations is consistent with respect to increasing information sets. That is, if $\mathcal{F}_t^* \subseteq \mathcal{F}_t$ for all t and $\{R_t\}_{t\in\mathbb{N}}$, $\{R_t^*\}_{t\in\mathbb{N}}$ are the optimal conditional forecasts with respect to their filtrations as defined at (2.8), then the internal procedure dominates the standard procedure, both conditionally and on average [Holzmann and Eulert (2014), Theorem 1]. The same is true if $\{R_t\}_{t\in\mathbb{N}}$ is $\mathcal{F}^*$-conditionally optimal and $\{R_t^*\}_{t\in\mathbb{N}}$ is just $\mathcal{F}^*$-predictable [Holzmann and Eulert (2014), Corollary 2]; see also Tsyplakov (2014).

Second, in the case $k = 1$, for most important functionals, including VaR and expectiles, strictly consistent scoring functions are *order sensitive* or *accuracy rewarding* in the following sense. Essentially, if $\Theta(\mathcal{L}(X)) < r < r^*$ or $r^* < r < \Theta(\mathcal{L}(X))$ for some random variable X, then

$$\mathbb{E}\big(S\big(\Theta\big(\mathcal{L}(X)\big), X\big)\big) < \mathbb{E}\big(S(r, X)\big) < \mathbb{E}\big(S\big(r^*, X\big)\big); \tag{2.16}$$

see Nau (1985), Lambert (2013) for details. Therefore, if the risk measure forecasts $\{R_t\}_{t\in\mathbb{N}}$ are always closer than $\{R_t^*\}_{t\in\mathbb{N}}$ to the optimal $\mathcal{F}^*$-conditional forecast, that is, $\Theta(\mathcal{L}(X_t|\mathcal{F}_t^*)) < R_t < R_t^*$ or $\Theta(\mathcal{L}(X_t|\mathcal{F}_t^*)) > R_t > R_t^*$ for all $t \in \mathbb{N}$ almost surely, then $\{R_t\}_{t\in\mathbb{N}}$ conditionally dominates $\{R_t^*\}_{t\in\mathbb{N}}$. There are different proposals for notions of order sensitivity in the case $k \ge 2$; see, for example, Lambert, Pennock and Shoham (2008), but the situation is less clear in this case.

The condition for conditional S-dominance in (2.15) can be formulated equivalently as

$$\mathbb{E}\big(\big(S(R_t, X_t) - S\big(R_t^*, X_t\big)\big)h_t\big) \le 0 \qquad \text{for all } h_t \ge 0,\ \mathcal{F}_{t-1}^*\text{-measureable}$$

for all $t \in \mathbb{N}$. It is tempting to work with a vector $\mathbf{h}_t$ of $\mathcal{F}^*$-predictable test functions in order to test for conditional S-dominance as suggested in the conditional calibration tests. However, we are interested in comparing the standard procedure to the internal procedure and reach a definite answer as to which one is to be preferred. If $\mathbb{E}((S(R_t, X_t) - S(R_t^*, X_t))\mathbf{h}_{t,i}) > 0$ but $\mathbb{E}((S(R_t, X_t) - S(R_t^*, X_t))\mathbf{h}_{t,j}) < 0$ for different components $\mathbf{h}_{t,i}, \mathbf{h}_{t,j}$ of the vector $\mathbf{h}_t$, no clear preference for either method can be given. Therefore, we do not pursue this approach further.

In comparative backtesting we are interested in the following null hypotheses:

H_0^- : The internal model predicts at least as well as the standard model,

H_0^+ : The internal model predicts at most as well as the standard model.

The null hypothesis H_0^- is analogous to the null hypothesis of a correct model and estimation procedure but now adapted to a comparative setting. As mentioned in the Introduction, considering a backtest as passed if the null hypothesis cannot be rejected is anti-conservative or aggressive in nature, and may therefore be problematic in regulatory practice. On the other hand, the null hypothesis H_0^+ is such that the comparative backtest is passed if we can reject H_0^+. This means that we can explicitly control the type I error of allowing an inferior internal model over an established standard model.

Let

$$\begin{aligned} \lambda^* &:= \limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\big(S(R_t, X_t) - S\big(R_t^*, X_t\big)\big), \\ \lambda_* &:= \liminf_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\big(S(R_t, X_t) - S\big(R_t^*, X_t\big)\big). \end{aligned} \tag{2.17}$$

It is clear that S-dominance on average implies $\lambda_* \le \lambda^* \le 0$. If the sequence of score differences $\{S(R_t, X_t) - S(R_t^*, X_t)\}_{t\in\mathbb{N}}$ is first-order stationary, then $\lambda^* = \lambda_*$, and $\lambda_* \le 0$ implies S-dominance on average. If λ^* in (2.17) is nonpositive, then the internal procedure is *at least as good* as the standard procedure, whereas the internal procedure *predicts at most as well* as the standard procedure if $\lambda_* \ge 0$. It may happen that λ_* and λ^* have different signs. Then we cannot order the two risk measurement procedures in terms of predictive performance. However, in finite samples this issue never occurs. Ordering risk measurement procedures is a compromise in the quest for conditional dominance. On the one hand, it is clearly a weaker notion than conditional dominance, but, on the other hand, in finite samples, it introduces a meaningful order on all risk measurement procedures given a sensible choice of the scoring function S; see Section 2.4.

Therefore, we reformulate our comparative backtesting hypotheses as

$$H_0^- : \quad \lambda^* \le 0,$$
$$H_0^+ : \quad \lambda_* \ge 0.$$

The test statistic

$$\Delta_n \overline{S} := \frac{1}{n} \sum_{t=1}^{n} \big(S(R_t, X_t) - S\big(R_t^*, X_t\big) \big)$$

for n large enough has expected value less or equal to zero under H_0^-, whereas under H_0^+ its expectation is non-negative. Tests of H_0^+ or H_0^- based on a suitably rescaled version of $\Delta_n \overline{S}$ are so-called *Diebold–Mariano tests*; see Diebold and Mariano (1995). Under certain mixing assumptions detailed in Giacomini and White (2006), Theorem 4,

$$\frac{\Delta_n \overline{S} - \mathbb{E}(\Delta_n \overline{S})}{\hat{\sigma}_n / \sqrt{n}}$$

is asymptotically standard normal with $\hat{\sigma}_n^2$ an HAC estimator [Andrews (1991)] of the asymptotic variance, $\sigma_n^2 = \operatorname{var}(\sqrt{n} \Delta_n \overline{S})$. Therefore, using the test statistic

$$T_4 = \frac{\Delta_n \overline{S}}{\hat{\sigma}_n / \sqrt{n}}, \tag{2.18}$$

we obtain an asymptotic level-η test of H_0^+ if we reject the null hypothesis when $\Phi(T_4) \le \eta$, and of H_0^- if we reject the null hypothesis when $1 - \Phi(T_4) \le \eta$.

Based on the outcome of the tests of H_0^+ and H_0^-, Fissler, Ziegel and Gneiting (2016) suggest the following three-zone approach. We fix a significance level $\eta \in (0, 1)$, for example, $\eta = 0.05$. If H_0^- is rejected at level η, then H_0^+ will not be rejected at level η. Similarly, if H_0^+ is rejected at level η, then H_0^- will not be rejected at level η. Therefore, we say that the internal procedure is in the red region; that is, it fails the comparative backtest if H_0^- is rejected. The internal procedure is in the green region; that is, it passes the backtest if H_0^+ is rejected. The internal procedure needs further investigation; that is, it falls in the yellow region if neither H_0^+ nor H_0^- can be rejected. For an illustration of these decisions, see Fissler, Ziegel and Gneiting (2016), Figure 1.

There is one important difference between the three-zone approach described in the Bank for International Settlements [(2013), pages 103–108] for traditional VaR backtests and the three-zone approach of Fissler, Ziegel and Gneiting (2016) described here. In the former approach, the zones arise from varying the confidence level of the hypothesis test, whereas in the latter approach the confidence level is fixed a priori, and the zones arise to separate cases where there is enough evidence to clearly decide for superiority of one procedure over the other in contrast to cases where there is no clear evidence.

2.4. *Choice of the scoring function.* Based on (2.2), (2.3) and (2.4), one has a large number of choices for strictly consistent scoring functions for VaR, expectiles and (VaR, ES). In the case of VaR_α, the standard choice is to take $G(r) = r$ in (2.2), leading to the classical asymmetric piecewise linear loss [see (2.19) below], also known as linlin, hinge, tick or pinball loss; see Koenker (2005) for its relevance in quantile regression. In the case of expectiles, one could argue that a natural choice is taking $\phi(r) = r^2$ in (2.3), which simplifies to the squared error function for the mean (up to equivalence). This is also the scoring function suggested by Newey and Powell (1987) for expectile regression. Consistent scoring functions for (VaR, ES) have only recently been discovered; see Acerbi and Szekely (2014), Fissler and Ziegel (2016). Therefore, there is no natural classical choice for the functions G_1, G_2 in (2.4).

A scoring function S is called *positive homogeneous* of degree b (or b-*homogeneous*) if for all $r = (r_1, \ldots, r_k)$ and all x

$$S(cr, cx) = c^b S(r, x) \qquad \text{for all } c > 0.$$

Efron (1991) argues that it is a crucial property of a scoring function to be positive homogeneous in estimation problems such as regression. Patton (2011) underlines the importance of positive homogeneity of the scoring function for forecast ranking. Positive homogeneous scoring functions are also favorable because they are so-called "unit consistent" [see, e.g., Acerbi and Szekely (2014)]; that is, if r and x are given in, say, U.S. dollars with $r = \$10$ and $s = \$5$, then, for a positive homogeneous scoring function S, the score $S(r, x) = S(\$10, \$5) = (\$)^b S(10, 5)$ will have unit (U.S. dollars)^b. In particular, changing the units, from, say, U.S. dollars to million U.S. dollars, will not change the ordering of forecasts assessed by this scoring function, and will thus also leave the results of comparative backtests unchanged. Concerning the choice of the degree b of homogeneity, Patton (2006) shows that, in the case of volatility forecasts, $b = 0$ requires weaker moment conditions than a larger choice of b for the validity of Diebold–Mariano tests which are used in comparative backtesting. Concerning the power of Diebold–Mariano tests, Patton and Sheppard (2009) find the best overall power for volatility forecasts for the choice $b = 0$.

Section C in the OS presents results which characterize positive homogeneous scoring functions for the risk measures that are of interest in this paper. Note that we only allow for predictions $r > 0$ or $r = (r_1, r_2)$ with $r_2 > 0$. As we are interested in risk measures for losses, this is not a real restriction; see also Section 3.2.

For some orders of homogeneity b, there is no strictly consistent scoring function for the risk measures of interest in this paper. In particular, the attractive choice $b = 0$ can often not be realized. However, for comparative backtesting we are not interested in absolute values of expected scores but only in *differences* of expected scores. Therefore, it is sufficient to have a scoring function such that the resulting score differences are homogeneous. Such homogeneous score differences of order

$b = 0$ exist for VaR, expectiles and (VaR, ES), as shown by the results delegated to the OS (Section C). Examples below list scoring functions which will be used subsequently in the simulation study and real data analysis.

EXAMPLE 4. For the comparative backtests for VaR that we investigate in Section 3.2, we consider the classical 1-homogeneous choice obtained by choosing $G(r) = r$ in (2.2), leading to the scoring function

$$S(r, x) = \big(1 - \alpha - \mathbb{1}\{x > r\}\big)r + \mathbb{1}\{x > r\}x. \tag{2.19}$$

Guided by the arguments given above, we alternatively consider the 0-homogeneous score differences by choosing $G(r) = \log r$, $r > 0$, which leads to the score

$$S(r, x) = \big(1 - \alpha - \mathbb{1}\{x > r\}\big)\log r + \mathbb{1}\{x > r\}\log x, \qquad r > 0. \tag{2.20}$$

EXAMPLE 5. The choice $\phi(r) = r^2$ in (2.3) leads to the strictly consistent scoring function

$$S(r, x) = -\mathbb{1}\{x > r\}(1 - 2\tau)(x - r)^2 + (1 - \tau)r(r - 2x) \tag{2.21}$$

for the τ-expectile e_τ. Besides this 2-homogeneous choice, in Section 3.2, we also investigate the 0-homogeneous alternative that arises by choosing $\phi(r) = -\log(r)$, $r > 0$, and hence we obtain the scoring function

$$S(r, x) = \mathbb{1}\{x > r\}(1 - 2\tau)\left(\log\frac{x}{r} + 1 - \frac{x}{r}\right) + (1 - \tau)\left(\log r - 1 + \frac{x}{r}\right). \tag{2.22}$$

EXAMPLE 6. For $(\mathrm{VaR}_\nu, \mathrm{ES}_\nu)$, we consider the $(1/2)$-homogeneous scoring function given by choosing $G_1(x) = 0$, $\mathcal{G}_2(x) = x^{1/2}$, $x > 0$ in (2.4) for comparative backtesting in Section 3.2. It is given by

$$S(r_1, r_2, x) = \mathbb{1}\{x > r_1\}\frac{x - r_1}{2\sqrt{r_2}} + (1 - \nu)\frac{r_1 + r_2}{2\sqrt{r_2}}. \tag{2.23}$$

As for the other risk measures, we also consider the 0-homogeneous alternative by choosing $G_1(x) = 0$, $\mathcal{G}_2(x) = \log x$, $x > 0$, which yields the scoring function

$$S(r_1, r_2, x) = \mathbb{1}\{x > r_1\}\frac{x - r_1}{r_2} + (1 - \nu)\left(\frac{r_1}{r_2} - 1 + \log(r_2)\right). \tag{2.24}$$

Acerbi and Szekely (2014) proposed a class of 2-homogeneous scoring functions for $(\mathrm{VaR}_\nu, \mathrm{ES}_\nu)$ depending on a parameter $W > 0$. It is strictly consistent when the class $\mathcal{P}$ of distributions is restricted to contain only distributions F with

$$\mathrm{ES}_\nu(F) < W\,\mathrm{VaR}_\nu(F).$$

In practice, it is generally not possible to say what magnitude of W is realistic to cover all possible applications. Therefore, we prefer to work with the homogeneous choices of strictly consistent scoring functions above and, more generally, of the form in Theorem C.3 of the OS.

3. Numerical studies.

3.1. *Forecasting of risk measures.* In this section we discuss a number of estimation procedures for producing conditional forecasts of the three risk measures discussed in this paper, namely, the VaR, expectile and ES. Owing to the widespread use of VaR in the banking sector, a great number of methods exist to produce its point forecasts; see, for example, Kuester, Mittnik and Paolella (2006) for an extensive review. In contrast, estimation and forecasting of expectiles in the risk measurement context is a relatively recent topic; see, for example, Kuan, Yeh and Hsu (2009). However, in many cases, similar methods as those used for VaR forecasting can be adopted for expectiles.

For illustrative purposes, we consider the following framework for forecasting of the risk measures. Suppose the series of negated log-returns $\{X_t\}_{t\in\mathbb{N}}$ can be modeled as

$$X_t = \mu_t + \sigma_t Z_t, \tag{3.1}$$

where $\{Z_t\}_{t\in\mathbb{N}}$ is a sequence of i.i.d. random variables with zero mean and unit variance, and μ_t and σ_t are measurable with respect to the sigma algebra $\mathcal{F}_{t-1}$, representing the information about the process $\{X_t\}_{t\in\mathbb{N}}$ available up to time $t-1$. In order to capture typical time dynamics of financial time series, one possibility is to assume that the conditional mean μ_t follows an ARMA process, while the condition variance σ_t^2 evolves according to a GARCH model specification.

Let ρ denote any of the three risk measures we consider. In the above setting, conditionally on the information up to time $t-1$, the one-step ahead forecast of ρ is

$$\rho\big(\mathcal{L}(X_t \mid \mathcal{F}_{t-1})\big) = \mu_t + \sigma_t \rho\big(\mathcal{L}(Z)\big), \tag{3.2}$$

where Z is used to denote a generic random variable with the same distribution as the Z_t's. Following McNeil and Frey (2000) and Diebold, Schuermann and Stroughair (2000), one can adopt a two-stage estimation procedure for the forecast $\rho(\mathcal{L}(X_t \mid \mathcal{F}_{t-1}))$. First μ_t and σ_t are estimated via the maximum likelihood procedure under a specific assumption[4] on the distribution of the innovations Z_t in (3.1). The second stage involves estimation of $\rho(\mathcal{L}(Z))$, the risk measure for i.i.d. sequence $\{Z_t\}_{t\in\mathbb{N}}$, based on the sample of standardized residuals

$$\big\{\hat{z}_t = (x_t - \hat{\mu}_t)/\hat{\sigma}_t\big\}. \tag{3.3}$$

[4]An alternative is to use the quasi-maximum likelihood estimation (MLE) procedure in which innovations Z_t are assumed to be standard normal. This is justified by a result in Bollerslev and Wooldridge (1992) stating that μ_t and σ_t would be consistently estimated even if the distribution of innovations is not normal, provided that the models for μ_t and σ_t are correctly specified. As pointed out in Kuester, Mittnik and Paolella (2006), the correct specification of dynamics for μ_t and σ_t may be difficult to fulfill in practice.

We consider the following three approaches to handle the second stage in the forecasting procedure: fully parametric (FP), filtered historical simulation (FHS), and a semiparametric estimation based on extreme value theory (EVT).

3.1.1. *Fully parametric estimation.* Under the fully parametric approach, a specific (parametric) model is assumed for the sequence of innovations $\{Z_t\}_{t\in\mathbb{N}}$. Examples of typically used probability distributions include the normal, Student's t and a skewed-t distribution [see, e.g., Fernández and Steel (1998)]. Parameters of the assumed distribution for Z_t's, denoted F_Z, can be estimated based on the standardized residuals $\{\hat{z}_t\}$ in (3.3) using, for example, the maximum likelihood method. If the model for Z_t's coincides with the one used to estimate the filter in the first stage, then no additional estimation is required at the second stage with all model parameters coming directly from the first stage estimation. The fitted distribution is used to compute the estimate of a given risk measure. In the case of $\mathrm{VaR}_\alpha(Z)$, this is given by the α-quantile, $\hat{F}_Z^{-1}(\alpha)$, whereas a τ-expectile $e_\tau(Z)$ can be computed as discussed in Section B.1 of the OS, where we give analytic expressions for expectiles of several commonly used distributions. Since we consider only continuous distributions F_Z, the ES can be computed as

$$\mathrm{ES}_\nu(Z) = \mathbb{E}\big(Z \mid Z \geq \mathrm{VaR}_\nu(Z)\big),$$

where we use numerical integration to evaluate the conditional expectation.

3.1.2. *Filtered historical simulation.* The method employs a nonparametric estimation of the risk measures based on the standardized residuals $\{\hat{z}_t\}$ in (3.3), which can be seen as representing a filtered time series; see, for example, Christoffersen (2003), Chapter 5.6. In particular, we draw a sample $\{\hat{z}_i^*; 1 \leq i \leq N\}$ of a large size N (e.g., $N = 10{,}000$) from $\{\hat{z}_t; 1 \leq t \leq n\}$ and then take the empirical estimate of a given risk functional as the estimate for $\rho(\mathcal{L}(Z))$. The empirical α-quantile gives the VaR estimate $\widehat{\mathrm{VaR}}_\alpha^{\mathrm{FHS}}(Z)$. The empirical τ-expectile $\hat{e}_\tau^{\mathrm{FHS}}(Z)$ is obtained using the least asymmetric weighted squares via iterative minimization of

$$\sum_{i=1}^{N} \omega_i(\tau)\big(\hat{z}_i^* - e_\tau\big)^2,$$

$$\omega_i(\tau) = \tau \mathbb{1}\big\{\hat{z}_i^* > e_\tau\big\} + (1-\tau)\mathbb{1}\big\{\hat{z}_i^* < e_\tau\big\} \qquad \text{with respect to } e_\tau.$$

The ES is estimated by the empirical version of the conditional expectation given that the residual exceeds the corresponding VaR estimate:

$$\widehat{\mathrm{ES}}_\nu^{\mathrm{FHS}}(Z) = \frac{1}{\#\{i : i = 1, \ldots, N, \hat{z}_i^* > \widehat{\mathrm{VaR}}_\alpha^{\mathrm{FHS}}(Z)\}} \sum_{i=1}^{N} \hat{z}_i^* \mathbb{1}\big\{\hat{z}_i^* > \widehat{\mathrm{VaR}}_\alpha^{\mathrm{FHS}}(Z)\big\}.$$

3.1.3. *EVT-based semiparametric estimation.* Risk is naturally associated with extremal events, and hence risk measure estimates rely on accurate estimation of a tail of the underlying distribution. However, inference about the distributional tails is notoriously difficult, as there are frequently not enough data points in the tail regions either to give a proper justification for a parametric model or to obtain reliable empirical estimates. Hence, unless a sufficiently long time series is available relative to the desired risk level for risk measure estimation, the two methods outlined in Sections 3.1.1 and 3.1.2 are unlikely to produce accurate forecasts. An alternative is to base estimation on asymptotic results of extreme value theory (EVT). For a detailed account, refer to, for example, Embrechts, Klüppelberg and Mikosch (1997).

The main premise is that, for a sufficiently high threshold u, conditional excesses of random variable Z satisfy

$$Z - u \mid Z > u \sim \mathrm{GP}(\beta_u, \xi), \tag{3.4}$$

where $\mathrm{GP}(\beta, \xi)$ denotes the generalized Pareto distribution with scale $\beta > 0$ and shape parameter $\xi \in \mathbb{R}$. It is common in applications to set the threshold at an upper order statistic, that is, $u = z_{(k+1)}$ for some $k < n$, where $z_{(1)} > z_{(2)} > \cdots > z_{(n)}$ are the decreasing order statistics of the sample $\{z_1, \ldots, z_n\}$ from F_Z. This leads to the following EVT-based estimates of $\mathrm{VaR}_\alpha(Z)$ and $\mathrm{ES}_\nu(Z)$ [see McNeil and Frey (2000)]:

$$\widehat{\mathrm{VaR}}_\alpha^{\mathrm{EVT}}(Z) = u + \frac{\hat\beta_u}{\hat\xi}\left(\left(\frac{k}{\alpha n}\right)^{\hat\xi} - 1\right), \qquad \hat\xi \neq 0, \tag{3.5}$$

and

$$\widehat{\mathrm{ES}}_\nu^{\mathrm{EVT}}(Z) = \widehat{\mathrm{VaR}}_\nu^{\mathrm{EVT}}(Z)\left(\frac{1}{1-\hat\xi} + \frac{\hat\beta - \hat\xi u}{(1-\hat\xi)\widehat{\mathrm{VaR}}_\nu^{\mathrm{EVT}}(Z)}\right), \tag{3.6}$$

with $(\hat\beta_u, \hat\xi)$ being parameter estimates of the GP distribution fitted to excesses over u. In the spirit of the above EVT-based estimators for VaR and ES, we derive an estimator for the τ-expectile. The details are provided in Section B.2 of the OS.

In the discussion above we assume that threshold u or, equivalently, k, the number of upper order statistics, is given so as to ensure adequacy of the approximation in (3.4). However, in practice, an accurate choice has to be made to balance the bias-variance trade-off, as a too large value of u increases variability of the parameter estimates of β_u and ξ, while insufficiently large u introduces the bias due to invalidity of (3.4). Various techniques have been proposed to assist with the choice of threshold such as graphical tools based on linearity of the mean excess function. As such methods require judgement at every time step at which conditional forecasts of risk measures are to be made, they are prohibitive for our purposes. Hence we adopt a pragmatic approach as in McNeil and Frey (2000), and take $k = 60$ in samples of size $n = 500$.

3.2. *Simulation study.* In practice, traditional backtesting is perhaps the most commonly used way to evaluate and subsequently choose among a number of competing forecasting procedures. While traditional backtesting is certainly suitable to capture some aspects of forecasting procedures, it does not provide information on the relative performance of different procedures with respect to the accuracy of forecasts, a seemingly natural criterion for a forecasting method. The aim of the present simulation study is to illustrate the use of the methodologies for traditional and comparative backtests discussed in the paper as well as to highlight the different messages delivered by the two types of backtests.

3.2.1. *Setup and forecasting methods.* The data $\{X_t\}_{t\in\mathbb{Z}}$ used for the analysis are generated from an AR(1)–GARCH(1, 1) process:

$$\begin{aligned} X_t &= \mu_t + \epsilon_t, \qquad \mu_t = -0.05 + 0.3X_{t-1}, \\ \epsilon_t &= \sigma_t Z_t, \qquad \sigma_t^2 = 0.01 + 0.1\epsilon_{t-1}^2 + 0.85\sigma_{t-1}^2, \end{aligned} \tag{3.7}$$

where innovations $\{Z_t\}_{t\in\mathbb{Z}}$ form a sequence of independent random variables with a common skewed-t distribution (see Example B.6 in the OS) with shape parameter $\nu = 5$ and skewness parameter $\gamma = 1.5$.

Quality of a forecasting procedure is determined by various factors. In a parametric or semiparametric set-up, potential model misspecification as well as estimation uncertainty in small samples can be detrimental for prediction. Nonparametric methods, while requiring no assumptions on the underlying model, are also subject to sampling variability and have strong limitations when dealing with extreme or tail events. The forecasting procedures we consider in the simulation study aim to cover a spectrum of models and estimation methods. We assume that the underlying process follows an AR(1)–GARCH(1, 1) dynamics with innovations $\{Z_t\}_{t\in\mathbb{Z}}$ coming from one of the following three distributions: the normal, the Student's t and the skewed-t distribution as in Example B.6 in the OS. We then consider the following estimation procedures:

- fully parametric estimation (Section 3.1.1) with the methods abbreviated as "n-FP", "t-FP" and "st-FP" under the assumption of normal, t and skewed-t distributed innovations, respectively;
- filtered historical simulation (Section 3.1.2) with the methods abbreviated as "n-FHS", "t-FHS" and "st-FHS";
- EVT-based estimation (Section 3.1.3) with the methods abbreviated as "n-EVT", "t-EVT" and "st-EVT".

In addition to the abovementioned methods, we supplement results with the optimal forecasts (abbreviated as "opt"), which uses the knowledge of the data-generating process.

Estimation is conducted using the moving window of size 500, and forecasts are evaluated based on the out-of-sample size of 5000 verifying observations. Section D of the OS provides an additional study in which only 250 verifying observations are used for backtesting.

The analysis is implemented using the open source software R [R Core Team (2015)]. The code is available at https://github.com/nnolde/Elicitability-and-Backtesting.

3.2.2. *Backtesting of risk measure forecasts*. Table 1 contains an overview of the one-step ahead forecasts obtained under the procedures described in the previous section. In particular, we report the average forecasts based on the series of moving estimation windows for each of the three considered risk measures, denoted $\overline{\mathrm{VaR}}_\alpha$, $\overline{e}_\tau$ and $\overline{\mathrm{ES}}_\nu$. The α levels for VaR are chosen in accordance with typical values used for internal risk management (such as $\alpha = 0.90$ and $\alpha = 0.95$) as well as the standard Basel VaR level $\alpha = 0.99$. For expectiles and ES, the levels are selected in such a way that the risk measure forecasts agree under the standard normal model.

In order to link to the previously used approaches to assess the quality of VaR forecasts (and to make comparisons between the methods), we computed the percentage of times the observations exceeded the VaR_α forecasts, commonly referred to as the percentage of violations. Based on the values reported under the column "% Viol." in Table 1, we observe that some of the misspecified models were actually able to hit nearly exactly the expected proportion of violations by matching the risk measure level $(1 - \alpha)$. This is the case, for instance, for "n-EVT" and "t-EVT" methods at $\alpha = 0.99$. Although large deviations from the risk measure confidence level do suggest substantial method deficiencies (as in the case of "n-FP" and "t-FP" methods), these values also highlight that the deviations from the $(1 - \alpha)$ level alone are unlikely to provide a good basis for differentiating the methods' performance in terms of prediction.

In addition to risk measure average forecasts, Table 1 also reports the average scores along with the corresponding method rankings using two different (consistent) scoring functions for each of the three considered risk measures. As the scoring functions we use require risk measure forecasts to be positive, we set the scores across all methods to zero in those few cases where forecasts are negative. Note that in the case of $(\mathrm{VaR}_\nu, \mathrm{ES}_\nu)$, only the forecasts for ES_ν are restricted to be positive.

The method rankings based on the average scores appear to be reasonable, and suggest some more general conclusions with respect to method selection on the basis of forecasting accuracy. Similar to the results of traditional backtesting, the numerical values in Table 1 provide further support to the observation that the choice of the likelihood model in fitting the AR(1)–GARCH(1, 1) filter has an appreciable influence on the accuracy of forecasts, perhaps more than previously thought

TABLE 1

Risk measure forecasts and method comparisons based on the sample average of consistent scoring functions in the simulation study; see Section 3.2 for details. The average scores $\overline{S}$ are divided by one minus the associated risk measure level to avoid very small values for presentation purposes. "% Viol." column shows the percentage of times observations exceeded the corresponding forecasts of VaR_α. The values in brackets indicate method ranks based on their average scores

	$\alpha = 0.90$				$\tau = 0.96561$			$\nu = 0.754$		
Method	$\overline{\mathrm{VaR}}_\alpha$	**% Viol.**	$\frac{1}{1-\alpha}\overline{S}$ **[eq. (2.19)]**	$\frac{1}{1-\alpha}\overline{S}$ **[eq. (2.20)]**	$\overline{e}_\tau$	$\frac{1}{1-\tau}\overline{S}$ **[eq. (2.21)]**	$\frac{1}{1-\tau}\overline{S}$ **[eq. (2.22)]**	$\overline{\mathrm{ES}}_\nu$	$\frac{1}{1-\nu}\overline{S}$ **[eq. (2.23)]**	$\frac{1}{1-\nu}\overline{S}$ **[eq. (2.24)]**
n-FP	0.440	9.4	0.7496 (9)	−0.4325 (7)	0.440	1.0149 (9)	−1.0526 (9)	0.440	0.6685 (10)	−0.8119 (9)
n-FHS	0.406	10.2	0.7484 (8)	−0.4288 (9)	0.542	1.0006 (7)	−1.3076 (7)	0.450	0.6626 (5)	−0.8361 (4)
n-EVT	0.406	10.2	0.7477 (7)	−0.4304 (8)	0.553	1.0039 (8)	−1.3188 (5)	0.449	0.6655 (9)	−0.8270 (8)
t-FP	0.348	12.2	0.7527 (10)	−0.3944 (10)	0.424	1.0200 (10)	−0.904 (10)	0.421	0.6645 (7)	−0.8040 (10)
t-FHS	0.413	10.0	0.7473 (6)	−0.4350 (5)	0.550	0.9899 (5)	−1.3055 (8)	0.456	0.6622 (4)	−0.8356 (5)
t-EVT	0.410	10.3	0.7471 (5)	−0.4329 (6)	0.562	0.9944 (6)	−1.3137 (6)	0.457	0.6654 (8)	−0.8289 (7)
st-FP	0.417	9.9	0.7442 (2)	−0.4391 (2)	0.559	0.9865 (4)	−1.3378 (3)	0.461	0.6606 (2)	−0.8460 (3)
st-FHS	0.412	10.1	0.7451 (4)	−0.4387 (3)	0.550	0.9808 (2)	−1.3342 (4)	0.455	0.6606 (3)	−0.8488 (2)
st-EVT	0.410	10.2	0.7449 (3)	−0.4363 (4)	0.561	0.9844 (3)	−1.3409 (2)	0.457	0.6642 (6)	−0.8350 (6)
opt	0.424	9.5	0.7431 (1)	−0.4454 (1)	0.565	0.9643 (1)	−1.4257 (1)	0.467	0.6575 (1)	−0.8704 (1)
	$\alpha = 0.95$				$\tau = 0.98761$			$\nu = 0.875$		
n-FP	0.586	5.9	0.9925 (8)	−0.1055 (9)	0.586	1.9845 (10)	−0.4650 (10)	0.587	0.8177 (10)	−0.3975 (10)
n-FHS	0.632	5.0	0.9910 (7)	−0.1123 (7)	0.801	1.8718 (7)	−0.8939 (5)	0.667	0.8121 (8)	−0.4261 (7)
n-EVT	0.628	5.1	0.9930 (9)	−0.1080 (8)	0.810	1.8756 (8)	−0.8935 (6)	0.670	0.8121 (7)	−0.4259 (8)
t-FP	0.518	7.3	1.0106 (10)	−0.0555 (10)	0.631	1.9008 (9)	−0.6419 (9)	0.716	0.8137 (9)	−0.4233 (9)
t-FHS	0.631	5.1	0.9902 (5)	−0.1148 (5)	0.822	1.8428 (5)	−0.8929 (7)	0.675	0.8112 (5)	−0.4292 (5)
t-EVT	0.630	5.1	0.9910 (6)	−0.1128 (6)	0.826	1.8506 (6)	−0.8885 (8)	0.677	0.8117 (6)	−0.4274 (6)
st-FP	0.639	4.9	0.9858 (2)	−0.1227 (2)	0.832	1.8313 (4)	−0.9156 (3)	0.688	0.8096 (3)	−0.4356 (3)
st-FHS	0.632	5.0	0.9887 (3)	−0.1161 (3)	0.821	1.8164 (2)	−0.9174 (2)	0.675	0.8096 (2)	−0.4357 (2)
st-EVT	0.630	5.1	0.9890 (4)	−0.1154 (4)	0.825	1.8221 (3)	−0.9153 (4)	0.677	0.8100 (4)	−0.4341 (4)
opt	0.649	4.7	0.9834 (1)	−0.1267 (1)	0.837	1.7481 (1)	−1.0189 (1)	0.696	0.8070 (1)	−0.4503 (1)

TABLE 1
(*Continued*)

	$\alpha = 0.99$				$\tau = 0.99855$			$\nu = 0.975$		
Method	$\overline{\text{VaR}}_\alpha$	**% Viol.**	$\frac{1}{1-\alpha}\overline{S}$ **[eq. (2.19)]**	$\frac{1}{1-\alpha}\overline{S}$ **[eq. (2.20)]**	$\overline{e}_\tau$	$\frac{1}{1-\tau}\overline{S}$ **[eq. (2.21)]**	$\frac{1}{1-\tau}\overline{S}$ **[eq. (2.22)]**	$\overline{\text{ES}}_\nu$	$\frac{1}{1-\nu}\overline{S}$ **[eq. (2.23)]**	$\frac{1}{1-\nu}\overline{S}$ **[eq. (2.24)]**
n-FP	0.859	2.5	1.8649 (10)	0.7041 (10)	0.859	8.4605 (10)	2.1097 (10)	0.863	1.1638 (10)	0.3969 (10)
n-FHS	1.193	1.1	1.7398 (8)	0.4992 (7)	1.492	6.1819 (7)	0.0652 (6)	1.218	1.1268 (8)	0.2453 (8)
n-EVT	1.189	1.0	1.7115 (5)	0.4801 (5)	1.480	6.1153 (5)	0.0651 (5)	1.243	1.1240 (7)	0.2381 (7)
t-FP	0.948	1.8	1.7605 (9)	0.5679 (9)	1.186	6.0364 (3)	0.2244 (9)	1.781	1.1472 (9)	0.2847 (9)
t-FHS	1.207	1.1	1.7392 (7)	0.5025 (8)	1.629	6.7232 (9)	0.0771 (8)	1.246	1.1205 (5)	0.2334 (6)
t-EVT	1.203	1.0	1.7064 (4)	0.4755 (4)	1.546	6.1387 (6)	0.0658 (7)	1.266	1.1208 (6)	0.2328 (5)
st-FP	1.214	0.9	1.6987 (3)	0.4734 (3)	1.583	5.9688 (2)	−0.0491 (2)	1.287	1.1156 (2)	0.2195 (2)
st-FHS	1.209	1.1	1.7339 (6)	0.4991 (6)	1.614	6.4895 (8)	0.0236 (3)	1.245	1.1161 (3)	0.2221 (4)
st-EVT	1.202	0.9	1.6929 (2)	0.4651 (2)	1.543	6.0779 (4)	0.0306 (4)	1.265	1.1164 (4)	0.2215 (3)
opt	1.227	0.9	1.6614 (1)	0.4369 (1)	1.574	4.9567 (1)	−0.3749 (1)	1.297	1.1066 (1)	0.1887 (1)

in the context of using the quasi-maximum-likelihood methods. Within each likelihood model, at lower levels for risk measure, fully parametric and FHS approaches tend to demonstrate better predictive performance, whereas at higher levels EVT-based methods seem to have an advantage, in particular, in the case of VaR. When the likelihood model is misspecified in fitting the AR(1)–GARCH(1, 1) filter, the nonparametric methods such as FHS and the semiparametric methods such as EVT-based estimation allow for greater flexibility to diminish the effects of model misspecification than the fully parametric approaches do. While in many cases rankings obtained from each pair of consistent scoring functions coincide, there also exist some discrepancies. This is not a surprise in the presence of misspecified models and estimation uncertainty as already pointed out by Patton (2014). For models for which the mean score is finite, the weak law of large numbers suggests convergence of the sample average (score) to the true mean (score) as the out-of-sample size tends to infinity. However, the convergence can be fairly slow. We found that in our simulation study the out-of-sample size of at least 1000 data points is necessary to achieve some stability in rankings. Hence, in finite sample situations, one has to be aware of the effects of sampling variability on the final rankings of the forecasting methods. Section D of the OS discusses results of a study where only 250 verifying observations were considered to perform backtesting. In small samples, results of both traditional and comparative backtesting may be greatly distorted by unrepresentative samples even when the underlying data-generating process is stationary.

Table 2 illustrates the traditional backtesting methodology presented in Section 2.2. Test statistics T_1 in (2.11) and T_2 in (2.12) are used, respectively, for two-sided and one-sided conditional calibration tests. The one-sided tests for VaR_α and τ-expectile are tests for super-calibration with p-values given by $\Phi(T_2)$. In the case of $(\mathrm{VaR}_\nu, \mathrm{ES}_\nu)$, we make use of Hommel's procedure [Hommel (1983)] with the adjusted p-values computed as $\tilde{\pi} = qC_q \min\{\pi_{(m)}/m; m = 1, 2\}$ and capped at one, where $\pi_m = 1 - \Phi(T_{2,m})$ for the one-sided tests of sub-calibration; see (2.13). (The classical Bonferroni multiple test procedure resulted in qualitatively similar conclusions.) For the simple conditional calibration tests, we set $\boldsymbol{h}_t = 1$. The test functions that were found to work well in this simulation study for general conditional calibration tests are

$$\mathbf{h}_t = \begin{cases} (1, r_t)' & \text{for } \mathrm{VaR}_\alpha, \\ \hat{\sigma}_t^{-1} & \text{for expectile } e_\tau, \\ \hat{\sigma}_t^{-1}\big((r_{2,t} - r_{1,t})/(1-\nu), 1\big) & \text{for } (\mathrm{VaR}_\nu, \mathrm{ES}_\nu) \end{cases} \tag{3.8}$$

in the case of two-sided tests, and

$$\mathbf{h}_t = \begin{cases} \big(1, |r_t|\big)' & \text{for } \mathrm{VaR}_\alpha, \\ \hat{\sigma}_t^{-1} & \text{for expectile } e_\tau, \\ \begin{pmatrix} 1 & |r_{1,t}| & 0 & 0 \\ 0 & 0 & 1 & \hat{\sigma}_t^{-1} \end{pmatrix}' & \text{for } (\mathrm{VaR}_\nu, \mathrm{ES}_\nu) \end{cases} \tag{3.9}$$

in the case of one-sided tests. The choice of test functions is important as it affects the properties of the test. For example, we found that inclusion of the lagged values of the identification function as in Example 1 resulted in tests which rejected all of the models including the optimal forecaster for $\mathrm{VaR}_{0.99}$ in the two-sided conditional calibration tests. A possible explanation for this phenomenon is that for a chosen test function the distribution of the test statistic becomes heavily skewed, making convergence to the asymptotic distribution slow. Another contributing factor, suggested by a referee, could be the instability of the $\hat{\Omega}^{-1}$ estimate in (2.11) due to high correlation of lagged values of the identification function. As discussed in Giacomini and White (2006), the choice of the test function with too few or too many components will also have direct implications on the power of the tests.

As expected, the numerical results in Table 2 show that the backtesting decisions based on the general conditional calibration tests are more conservative in comparison to the corresponding simple conditional calibration tests, subject to a sensible choice of the test function. This is particularly visible for one-dimensional risk measures (VaR and expectiles) when performing the two-sided tests. The two-sided conditional calibration tests for these two risk measures suggest the importance of the correct specification of the likelihood used in fitting the AR(1)–GARCH(1, 1) filter. The entirely parametric methods with misspecified models (here "n-FP" and "t-FP") fail traditional backtests even when testing for simple conditional calibration (with the exception of $\mathrm{VaR}_{0.90}$). The general conditional tests are able to pick up the misspecified likelihoods at least in some instances, for example, when forecasting $\mathrm{VaR}_{0.90}$ and using the (symmetric) t distribution instead of the true asymmetric underlying model, and similarly for τ-expectiles with $\tau = 0.96561$ and $\tau = 0.98761$. The general conditional two-sided calibration tests also detect the differences in the second stage of risk measure forecasting when different methods are applied to filtered series of innovations. For instance, at the highest risk measure levels, the EVT-based methods tend to pass the conditional backtests in contrast to their empirical and in some cases even parametric (correctly specified) counterparts; see panels for $\mathrm{VaR}_{0.99}$ and 0.99855-expectile. This is true even under a misspecified likelihood model in the AR(1)–GARCH(1, 1) filter.

We also note that the tests for one-dimensional risk measures appear to have better power properties than the tests for the two-dimensional risk measure, $(\mathrm{VaR}_\nu, \mathrm{ES}_\nu)$, although a more thorough investigation into finite sample properties of these tests would be necessary to draw more definitive conclusions. It can also be observed that the one-sided tests are less conclusive than their two-sided analogues. This is perhaps not a surprise, as it may well happen that a method is not good at predicting the risk measure but gives a correct bound, and thus should not be rejected by a one-sided calibration test.

Finally, Figures 1–3 display the traffic light matrices for the three risk measures and two forms of consistent scoring functions for each. These plots complement

TABLE 2

P-values for traditional backtests in the simulation study; see Section 3.2 for details. The one-sided tests for VaR_α *and* τ*-expectile are tests of super-calibration, and of sub-calibration for* $(\mathrm{VaR}_\nu, \mathrm{ES}_\nu)$*. The test functions used in general conditional calibration tests are given in* (3.8) *and* (3.9)*. Values in boldface are significant at the* 5% *level*

	$\alpha = 0.90$ VaR_α				$\tau = 0.96561$ τ-expectile				$\nu = 0.754$ $(\mathrm{VaR}_\nu, \mathrm{ES}_\nu)$			
	two-sided		one-sided		two-sided		one-sided		two-sided		one-sided	
Method	**simple**	**general**	**simple**	**general**	**simple**	**general**	**simple**	**general**	**simple**	**general**	**simple**	**general**
n-FP	0.146	**0.018**	0.927	1.000	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
n-FHS	0.576	0.058	0.288	0.863	0.887	**0.048**	0.443	0.193	0.881	0.184	0.712	0.744
n-EVT	0.608	0.056	0.304	0.911	0.684	**0.042**	0.658	0.364	0.754	0.672	1.000	0.629
t-FP	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.086	**0.006**	**0.041**	**0.011**
t-FHS	0.962	**0.006**	0.481	1.000	0.728	**0.030**	0.636	0.330	0.936	0.512	0.960	0.256
t-EVT	0.514	**0.011**	0.257	0.772	0.360	**0.023**	0.820	0.542	0.880	0.475	0.815	**0.008**
st-FP	0.740	0.090	0.630	1.000	0.429	0.084	0.786	0.546	0.569	0.824	1.000	0.991
st-FHS	0.851	0.091	0.425	1.000	0.708	0.123	0.646	0.400	0.909	0.796	0.956	0.744
st-EVT	0.674	0.066	0.337	1.000	0.377	0.098	0.812	0.596	0.935	0.706	0.851	**0.032**
opt	0.228	0.294	0.886	1.000	0.234	0.458	0.883	0.850	0.401	0.337	0.732	1.000
	$\alpha = 0.95$				$\tau = 0.98761$				$\nu = 0.875$			
n-FP	**0.006**	**0.004**	**0.003**	**0.009**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
n-FHS	0.948	**0.042**	0.526	1.000	0.702	0.067	0.351	0.158	0.912	0.349	0.997	0.609
n-EVT	0.797	0.075	0.398	1.000	0.868	0.062	0.434	0.208	0.720	0.549	1.000	0.762
t-FP	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	1.000	1.000
t-FHS	0.700	0.053	0.350	1.000	0.793	**0.027**	0.603	0.325	0.951	0.492	0.864	0.368
t-EVT	0.654	0.106	0.327	0.981	0.713	**0.033**	0.643	0.363	0.699	0.771	1.000	0.845
st-FP	0.794	0.261	0.603	1.000	0.568	0.066	0.716	0.467	0.655	0.898	0.907	0.249
st-FHS	0.897	0.111	0.449	1.000	0.729	0.073	0.635	0.393	0.908	0.690	0.904	0.875
st-EVT	0.797	0.180	0.398	1.000	0.643	0.077	0.679	0.435	0.599	0.968	1.000	1.000
opt	0.284	0.552	0.858	1.000	0.315	0.523	0.843	0.798	0.311	0.624	0.263	0.194

TABLE 2
(*Continued*)

	$\alpha = 0.99$ VaR_α				$\tau = 0.99855$ τ-expectile				$\nu = 0.975$ $(\text{VaR}_\nu, \text{ES}_\nu)$			
	two-sided		**one-sided**		**two-sided**		**one-sided**		**two-sided**		**one-sided**	
Method	**simple**	**general**	**simple**	**general**	**simple**	**general**	**simple**	**general**	**simple**	**general**	**simple**	**general**
n-FP	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
n-FHS	0.420	**0.007**	0.210	0.630	0.377	**0.045**	0.188	0.100	0.653	0.231	0.549	0.538
n-EVT	1.000	0.186	0.500	1.000	0.300	0.080	0.150	0.085	0.886	0.226	0.804	0.577
t-FP	**0.000**	**0.000**	**0.000**	**0.000**	**0.003**	**0.010**	**0.002**	**0.001**	**0.000**	**0.000**	1.000	1.000
t-FHS	0.679	**0.029**	0.339	1.000	0.783	**0.013**	0.391	0.212	0.697	0.717	1.000	1.000
t-EVT	0.888	0.140	0.444	1.000	0.509	0.067	0.254	0.145	0.995	0.498	0.807	1.000
st-FP	0.454	0.221	0.773	1.000	0.601	**0.048**	0.301	0.169	0.695	0.419	0.597	0.511
st-FHS	0.584	**0.018**	0.292	0.876	0.826	**0.026**	0.413	0.238	0.843	0.758	1.000	1.000
st-EVT	0.554	0.270	0.723	1.000	0.552	0.087	0.276	0.162	0.962	0.564	0.868	1.000
opt	0.364	0.576	0.818	1.000	0.825	0.491	0.588	0.513	0.131	0.571	0.073	0.101

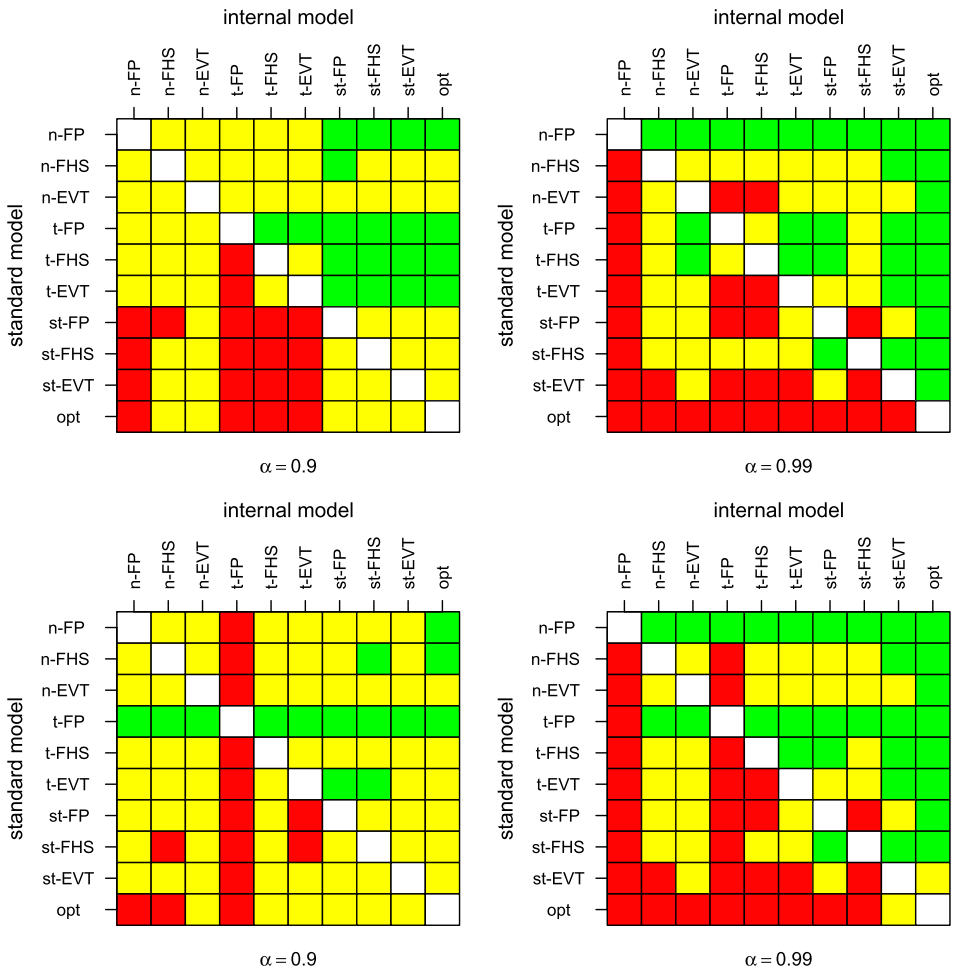

FIG. 1. *Traffic light matrices for* VaR_α *forecasts at the test confidence level* $\eta = 0.05$. *The top and bottom rows are based on the scoring functions in* (2.19) *and* (2.20), *respectively.*

the method rankings on the basis of just the average scores with the tests of predictive ability at the test level $\eta = 5\%$. Along the vertical axis we consider hypothetical "standard" models with the investigated "internal" models displayed along the horizontal axis. The red and green cells correspond to situations in which the comparative backtest is failed or passed, while yellow cells indicate cases where no conclusive evidence is available to pass or fail the comparative backtest. The rows in each figure correspond to different scoring functions used to compare the methods.

Inconclusive traffic light matrices can result if all methods are performing reasonably well or if the chosen scoring function has poor discrimination ability. In the case of VaR, as the discrimination ability of both chosen scoring functions

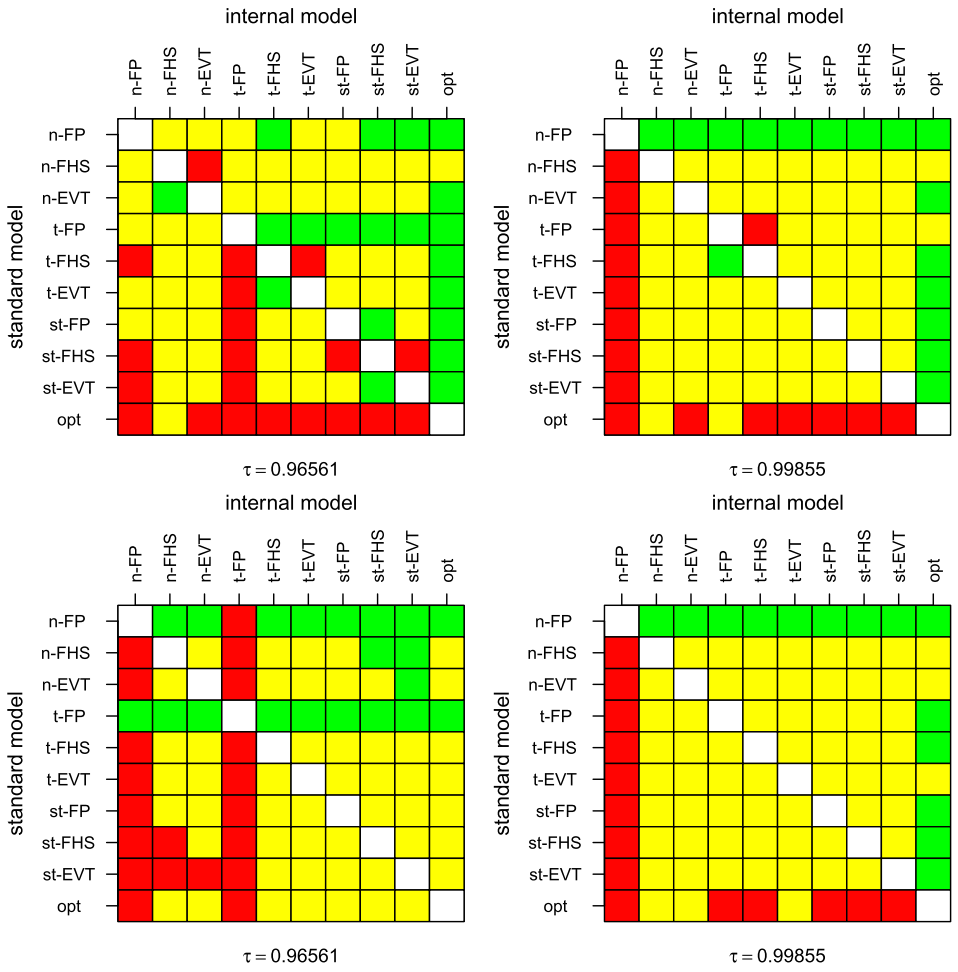

FIG. 2. *Traffic light matrices for τ-expectile forecasts at the test confidence level $\eta = 0.05$. The top and bottom rows are based on the scoring functions in* (2.21) *and* (2.22)*, respectively.*

seems good at level $\alpha = 0.99$, it is likely that at $\alpha = 0.90$ several models show a reasonable predictive ability. This is in line with the largely inconclusive traditional backtests at level $\alpha = 0.90$. At $\alpha = 0.90$, the scoring function in (2.19) is better at identifying models with the correctly specified likelihood than the scoring function in (2.20), for which with just a few exceptions only the "t-FP" method fails the comparative backtests as an internal method against all the other possible standard methods. At $\alpha = 0.99$, the two scoring functions result in a good agreement with "n-FP" being the worst forecaster (i.e., failing the comparative backtests against all the other methods), the optimal method passing comparative backtests against all other methods [the exception is "st-EVT" under the scoring function in (2.20)].

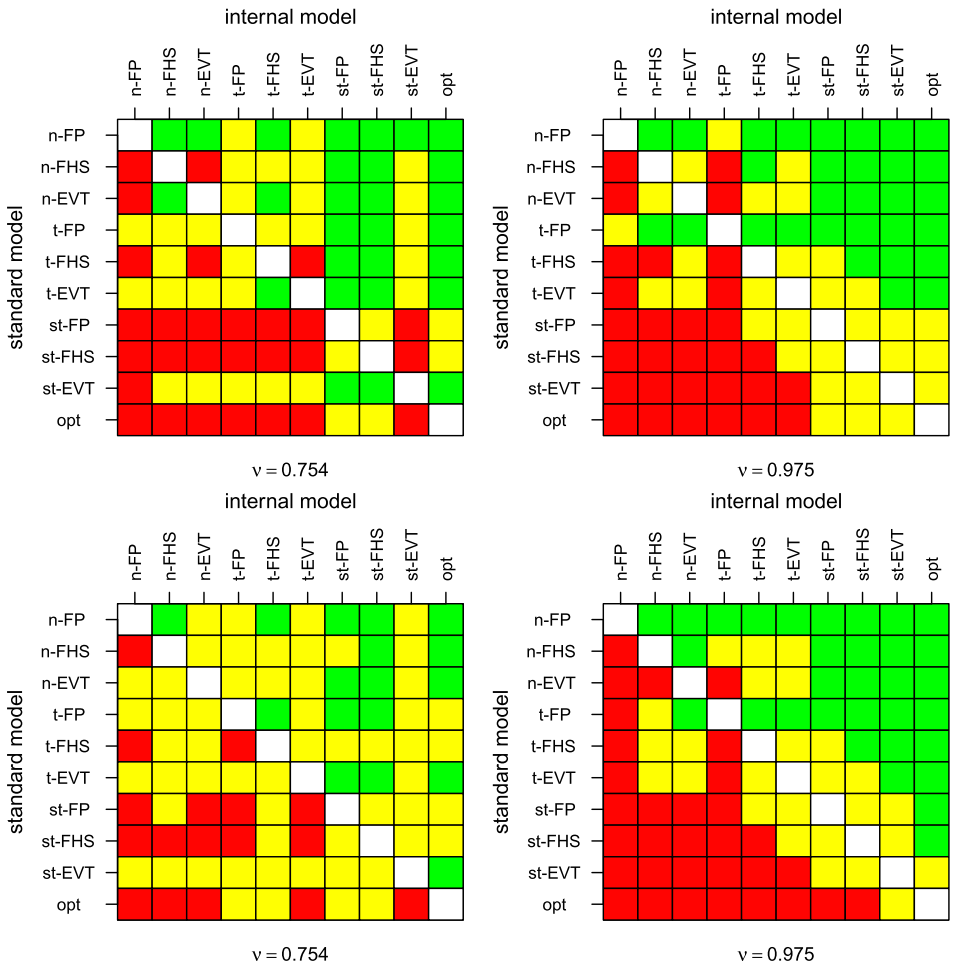

FIG. 3. *Traffic light matrices for* $(\mathrm{VaR}_\nu, \mathrm{ES}_\nu)$ *forecasts at the test confidence level* $\eta = 0.05$. *The top and bottom rows are based on the scoring functions in* (2.23) *and* (2.24), *respectively.*

The situation is less clear for the τ-expectile. At level $\tau = 0.96561$, the "n-FP" method fails the comparative backtest against most of the other methods under both scoring functions; the use of the scoring function in (2.22) suggests failing the "t-FP" method as well. The "st-EVT" method would pass the comparative backtest against the models with the normal likelihood and "t-FP." At level $\tau = 0.99855$, both scoring functions do not discriminate the methods much except for flagging the optimal forecaster as better than most other methods and failing the "n-FP" method. Expectiles have been used much less as a risk measure, and it may be possible that the present methods are indeed suboptimal for expectile prediction at high levels. Again, this is in line with the results of the traditional backtests, in particular, the conditional two-sided tests.

For $(\mathrm{VaR}_\nu, \mathrm{ES}_\nu)$, the large number of conclusive comparative backtesting results indicates that we can discriminate well between methods, and, as in the case of VaR, it appears less important which method to use at a lower level than at a higher level. In particular, we once again see that the methods with the correctly specified likelihood show superior predictive performance. According to the scoring function in (2.23), the "st-EVT" method fails the comparative backtest against its parametric and nonparametric counterparts "st-FP" and "st-FHS" at lower levels of ν. No definitive conclusions with respect to these models can be drawn at $\nu = 0.975$.

3.3. *Data analysis*: *NASDAQ composite index*. We have fitted an AR(1)–GARCH(1, 1) model to the negated log-returns of the NASDAQ Composite index using a moving estimation window of 500 data points. The time series we consider is from Feb. 8, 1971 until May 18, 2016, which gives us an out-of-sample size $n = 10{,}920$ to perform backtesting. The data is publicly available and has been downloaded from http://finance.yahoo.com. (The code and the data used are posted at https://github.com/nnolde/Elicitability-and-Backtesting.) While for illustrative purposes we used the entire time series available to us at the time of manuscript preparation, we do note that results are subject to sampling variability, especially if only a small out-of-sample size is available to perform backtesting. Please refer to Section D of the OS for further discussion of this issue.

Table 3 summarizes results of traditional and comparative backtesting for six forecasting methods (refer to Section 3.2 for details on these methods) and, as before, for the three risk measures [VaR, expectile and the (VaR, ES) pair] at their standard Basel levels.

In the case of $\mathrm{VaR}_{0.99}$, the traditional backtests based on the two-sided simple conditional calibration tests are passed only under the n-EVT and st-EVT methods. So here the choice of the likelihood function in fitting the AR(1)–GARCH(1, 1) filter seems to have a lower impact than the choice of the method at the second stage of forecasting applied to the fitted residuals. At this relatively high risk measure level, the EVT-based methods outperform their other competitors based on both the traditional backtests and the average scores. It should also be noted that the two scoring functions have led to the same rankings of the forecasting procedures. The fully parametric methods (n-FP and st-FP) show the worst performance in terms of their predictive ability. n-FP falls into the red region against all other methods, whereas st-FP fails against the EVT methods and cannot win against the FHS methods; see the traffic light matrices in Figure 4 (top row).

On the other hand, for the 0.99855-expectile, the tests of simple conditional calibration are rejected (at 5% level) for all the methods that use the normal likelihood. Those methods that use the skewed-t likelihood also tend to rank higher, although, in terms of significance, most methods fall into the yellow region (apart from the n-FP method). The ranking of forecasts is different for the two scoring functions

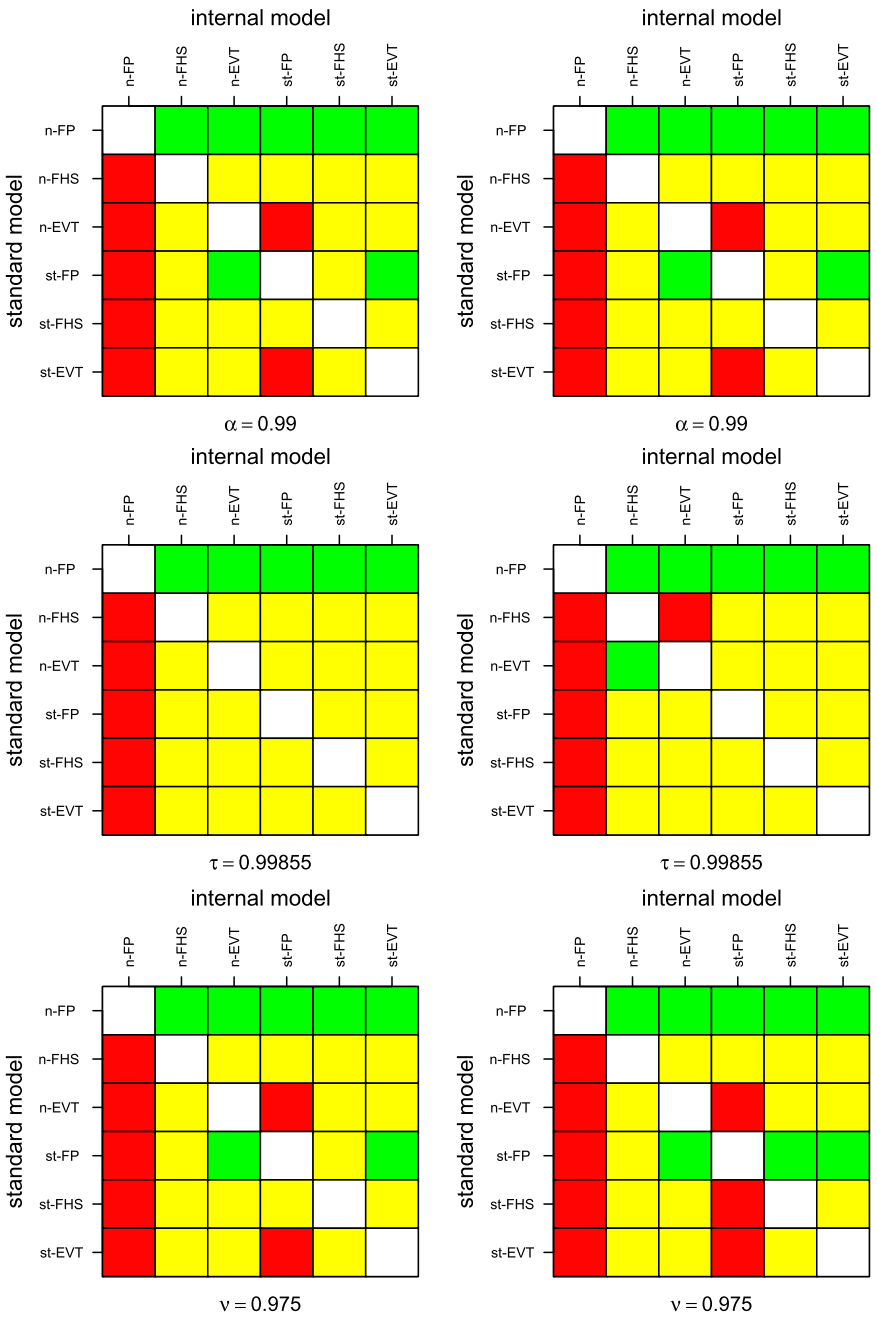

FIG. 4. *Traffic light matrices for* VaR_α *(top row) based on scoring functions in* (2.19) *(left) and* (2.20) *(right), for* τ*-expectile (middle row) based on scoring functions in* (2.21) *(left) and* (2.22) *(right), and for* $(\mathrm{VaR}_\nu, \mathrm{ES}_\nu)$ *(bottom row) based on scoring functions in* (2.23) *(left) and* (2.24) *(right) at the test confidence level* $\eta = 0.05$ *for the data analysis in Section* 3.3.

TABLE 3

Summary of traditional and comparative backtesting based on the negated log-returns on the NASDAQ Composite index with an AR(1)–GARCH(1, 1) *filter fitted over a moving estimation window of* 500 *observations and the out-of-sample size of* $n = 10{,}920$; *refer to Section* 3.3 *for details. The second column reports the average risk measure forecasts. "% Viol." gives the percentage of* $\mathrm{VaR}_{0.99}$ *forecast exceedances. The simple CCT and general CCT columns contain the p-values for two-sided simple and general conditional calibration tests, respectively. The final two columns show the average scores, scaled by one minus the risk measure confidence level for presentation purposes, based on the specified scoring functions along with the corresponding method ranks (in brackets)*

Method	$\overline{\mathrm{VaR}}_{0.99}$	**% Viol.**	**simple CCT**	**general CCT**	$\overline{S}$ **[eq. (2.19)]**	$\overline{S}$ **[eq. (2.20)]**
n-FP	2.363	2.3	**0.000**	**0.000**	3.8497 (6)	1.3017 (6)
n-FHS	2.758	1.3	**0.017**	**0.028**	3.5842 (3)	1.1604 (3)
n-EVT	2.774	1.2	0.112	0.152	3.5675 (2)	1.1550 (2)
st-FP	2.739	1.3	**0.004**	**0.012**	3.5976 (5)	1.1669 (5)
st-FHS	2.785	1.2	**0.046**	0.108	3.5904 (4)	1.1609 (4)
st-EVT	2.811	1.1	0.181	0.290	3.5654 (1)	1.1517 (1)
	$\overline{e}_{0.99855}$		**simple CCT**	**general CCT**	$\overline{S}$ **[eq. (2.21)]**	$\overline{S}$ **[eq. (2.22)]**
n-FP	2.363		**0.000**	**0.000**	25.9030 (6)	0.9660 (6)
n-FHS	2.986		**0.049**	**0.002**	19.7333 (2)	0.2933 (4)
n-EVT	2.966		**0.023**	**0.001**	19.8196 (5)	0.3084 (5)
st-FP	3.041		0.163	**0.011**	19.8159 (4)	0.2509 (1)
st-FHS	3.078		0.227	**0.011**	19.7533 (3)	0.2589 (2)
st-EVT	3.037		0.112	**0.006**	19.6963 (1)	0.2687 (3)
	$\overline{\mathrm{ES}}_{0.975}$		**simple CCT**	**general CCT**	$\overline{S}$ **[eq. (2.23)]**	$\overline{S}$ **[eq. (2.24)]**
n-FP	2.375		**0.000**	**0.000**	1.7020 (6)	1.0492 (6)
n-FHS	2.777		**0.022**	**0.035**	1.6587 (4)	0.9637 (4)
n-EVT	2.813		0.261	**0.015**	1.6560 (1)	0.9607 (2)
st-FP	2.810		**0.001**	0.248	1.6622 (5)	0.9691 (5)
st-FHS	2.816		0.139	0.067	1.6582 (3)	0.9617 (3)
st-EVT	2.857		0.327	0.117	1.6563 (2)	0.9597 (1)

used. The 0-homogeneous choice at (2.22) clearly ranks the methods using the normal likelihood lower than those using the skewed-t likelihood in agreement with the results of the simple conditional calibration tests, which is an argument in favor of using (2.22) rather than (2.21).

For both $\mathrm{VaR}_{0.99}$ and 0.99855-expectile, the conditional calibration tests with the test functions as in the simulation study lead to the failure of the corresponding traditional backtest; see Table 3 for the expectile. This may seem overly conservative for practical purposes, and suggests either reexamining suitability of the GARCH-type filter for these data or the use of a more appropriate test function. For $\mathrm{VaR}_{0.99}$, we performed the conditional calibration tests also with the test function

$\mathbf{h}_t = (1, V(r_{t-1}, x_{t-1}))'$ (see Example 1), and the resulting p-values are reported in Table 3. They lead to conclusions similar to those based on the simple conditional calibration tests. This example underlines the importance of further studies on appropriate choices of test functions.

The results for $(\mathrm{VaR}_\nu, \mathrm{ES}_\nu)$ with $\nu = 0.975$ suggest better performance when a more flexible model such as the skewed-t is used to fit the AR(1)–GARCH(1, 1) filter, although the use of EVT-based methods has a potential to compensate for likelihood misspecifications. Again, fully parametric methods (n-FP and st-FP) fall into the red region in the comparative backtests against most of the other more flexible alternatives; see the bottom panels in Table 3 and Figure 4. The outcomes show one interesting aspect which is not in contradiction with the theory but may be puzzling and merit further investigation in future studies: The conditional calibration test rejects all methods using a normal likelihood, but the scoring functions rank the n-EVT method as the best or second best performing method. It seems that the test function used in the conditional calibration test is sensitive to the likelihood function used in fitting the AR(1)–GARCH(1, 1) filter, whereas the scoring functions are more sensitive to the method at the second stage giving preference to the EVT methods.

4. Discussion. In the paper we have discussed two approaches to backtesting risk measure forecasts. We differentiate between traditional backtesting, which gives a "yes" or "no" answer to the question of whether a method is acceptable or not, and comparative backtesting, specifically aimed at comparing the predictive performance of different forecasting methods. In general, there appears to be a need for both traditional and comparative backtesting methodologies. The former poses a requirement of identifiability on the risk measure functional, and serves the purpose of categorizing methods based on whether the backtest is passed or not, albeit with a somewhat limited ability to fail misspecified models. However, traditional backtesting does not provide a statistically justifiable basis for method comparisons often sought when assessing the performance of, say, a newly proposed forecasting procedure against an existing one or when defending an internal procedure against some standard procedure. Comparative backtesting provides a methodology to serve exactly these purposes. For methods that are deemed acceptable under a traditional backtest, comparative backtesting allows to rank methods according to their predictive performance based on a chosen consistent scoring function, provided that the risk measure under consideration is an elicitable functional.

Traditional backtesting, which we formalize in the form of conditional calibration tests, provides a unifying framework for currently available backtests of risk measures. To assess performance of different calibration tests in a controlled environment, a simulation study was conducted. It emerged that in fact many methods based on misspecified models may pass traditional backtests. And while the outcome of the backtest is the same in all such cases (a pass), differences in risk

measure forecasts under different methods will ultimately lead to different capital requirements. One practical implication of this is that such backtests may create a wrong incentive of minimizing the capital, subject to passing the backtest, rather than aiming for a more accurate forecasting method. From the simulation study, we have also seen that general conditional calibration tests have a slightly better ability at detecting methods with misspecified models in comparison to the corresponding simple conditional calibration tests, with the latter being able to flag only the most under-performing methods. However, for the real data, often, simple and general conditional calibration tests produced similar results, suggesting that in practice the use of simple conditional calibration tests may suffice. General conditional calibration tests offer a more refined alternative, but require the choice of a test function. Further research is necessary to gain more insight into the choice of the test function for different risk measures and how this choice affects the outcomes of the tests.

In light of the above mentioned limitations of traditional backtests, regulators may additionally apply a comparative backtest in cases where a traditional backtest is passed. This necessitates a standard model against which the bank's internal model is to be tested. Such a standard model should not be confused with the standardized approaches currently used by regulators for trading book risk management of banks that either are not able to go for the (internal) model-based approach or do not pass the regulatory backtesting. These standardized approaches do not produce risk measure forecasts, and hence could not be incorporated into the comparative backtesting framework. However, comparative backtests will create the correct incentive for the banks to develop risk measure forecasting methods that aim for accuracy of forecasts, and hence can adequately quantify the risks. If the Basel committee were to introduce comparative backtesting, a foresting method to serve as the "standard model" should be chosen among flexible methods that have low model risk and are known to do well under a fairly broad range of circumstances. One such possibility could be the filtered historical simulation with a GARCH filter fitted using a flexible likelihood model such as the skewed-t in our numerical examples.

In summary, our recommendation to the Basel committee would be to adopt a two-stage backtesting framework. At stage I, a calibration test is applied in line with the current practice. In terms of implementation, the easiest option is to use the two-sided simple conditional calibration test. Conditionally on passing the stage I test, stage II will then assess the bank's "internal model" against the regulator's "standard model" via a comparative backtest. From the regulatory point of view, the statistical significance of the comparative backtests can be nicely summarized by means of traffic light matrices highlighting which methods pass or fail against a standard procedure, and when not enough evidence is available to make a conclusive statement. Provided that the regulatory risk measure is elicitable, comparative backtests require a choice of a consistent scoring function for that risk measure. In the case of backtesting ES, the current regulatory risk measure

for banks' trading books, the 0-homogeneous scoring function in equation (2.24) would be a reasonable choice, as it is unit consistent and has milder moment restrictions on the underlying stochastic process than other positive homogeneous alternatives. Additionally, based on the data analysis, it yields results in rankings which are in better agreement with the outcomes of the calibration tests and leads to slightly more conclusive results in terms of the traffic light matrix entries versus the considered $1/2$-homogeneous alternative.

It is worth noting that the comparative backtesting methodology can also be used by financial institutions internally to select better performing methods among competing alternatives. The same would apply to academic literature seeking to compare different forecasting methods, with the comparison done on the basis of forecast accuracy, in addition to calibration.

There are still many open problems and follow-up questions that require further investigation to create a fuller understanding of the usability of the presented backtesting methodologies. In the context of traditional backtesting, we found conditional calibration tests to be better at detecting model misspecifications. However, these conditional tests require the user to choose a set of test functions. An exploration of potential test function choices and their influence on finite sample properties of the tests in a broader context than covered in our simulation study would be beneficial to guide practical applicability of these backtests. A choice problem also arises in the context of comparative backtesting where it is possible to make use of any member of the family of consistent scoring functions for a given risk measure functional. Here, different aspects of the resulting backtests can be assessed. One particular aspect to consider is the existence of the mean score (or difference in scores) for the underlying process. Financial time series tend to have fairly heavy tails and this would place restrictions on the choice of a suitable scoring function. From this perspective, the proposed scoring functions with 0-homogeneous score differences allow to study heavier-tailed processes than the b-homogeneous choices (with $b > 0$). Finally, we have not explored the potentially promising possibility of using conditional comparative backtests. There are many open questions on how they should be formulated and implemented to be informative in practice.

Some of the risk measures used in practice are in fact nonelicitable. A prominent example here is the ES. In such cases the notion of joint elicitability may open the door to the ability to conduct backtests, in this case for multivariate risk measure functionals. We have explored the joint elicitability of VaR and ES, and, on the basis of our simulation study, the backtesting results show a good ability to identify and differentiate among methods relying on correct and misspecified model formulations. However, further research is needed to provide a clearer interpretation of both traditional and comparative backtests. For example, in the case of the pair (VaR, ES), the question would be whether it is a poor forecasting of VaR or ES or both that caused a (traditional or comparative) backtest to fail.

Acknowledgments. We are grateful to the Editor for his supportive and constructive guidance on improving the manuscript as well as to the five referees for providing many useful comments and suggestions. We would also like to thank Professor Paul Embrechts for a number of inspiring discussions, as well as RiskLab at ETH Zurich for its hospitality when we began working on this project.

SUPPLEMENTARY MATERIAL

Supplementary material for article "Elicitability and backtesting: Perspectives for banking regulation" (DOI: 10.1214/17-AOAS1041SUPP; .pdf). We elaborate on some of the points made in the main article as well as provide technical details and proofs of several results.

REFERENCES

ACERBI, C. and SZEKELY, B. (2014). Backtesting expected shortfall. *Risk Mag*. December 76–81.

ANDREWS, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59** 817–858. MR1106513

BANK FOR INTERNATIONAL SETTLEMENTS (2013). Consultative document: Fundamental review of the trading book: A revised marked risk framework. Available at http://www.bis.org/publ/bcbs265.pdf.

BANK FOR INTERNATIONAL SETTLEMENTS (2014). Consultative document: Fundamental review of the trading book: Outstanding issues. Available at http://www.bis.org/bcbs/publ/d305.pdf.

BELLINI, F. and BIGNOZZI, V. (2015). On elicitable risk measures. *Quant. Finance* **15** 725–733. MR3334566

BELLINI, F. and DI BERNARDINO, E. (2017). Risk management with expectiles. *Eur. J. Finance* **23** 487–506.

BELLINI, F., KLAR, B., MÜLLER, A. and GIANIN, E. R. (2014). Generalized quantiles as risk measures. *Insurance Math. Econom.* **54** 41–48. MR3145849

BOLLERSLEV, T. and WOOLDRIDGE, J. M. (1992). Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. *Econometric Rev.* **11** 143–172. MR1185178

CHRISTOFFERSEN, P. F. (1998). Evaluating interval forecasts. *Internat. Econom. Rev.* **39** 841–862. MR1661906

CHRISTOFFERSEN, P. (2003). *Elements of Financial Risk Management*. Academic Press.

CONT, R., DEGUEST, R. and SCANDOLO, G. (2010). Robustness and sensitivity analysis of risk measurement procedures. *Quant. Finance* **10** 593–606. MR2676786

COSTANZINO, N. and CURRAN, M. (2015). Backtesting general spectral risk measures with application to expected shortfall. *Journal of Risk Model Validation* **9** 21–33.

DAVIS, M. H. A. (2016). Verification of internal risk measure estimates. *Stat. Risk Model.* **33** 67–93. MR3574946

DELBAEN, F., BELLINI, F., BIGNOZZI, V. and ZIEGEL, J. F. (2016). Risk measures with the CxLS property. *Finance Stoch.* **20** 433–453. MR3479327

DIEBOLD, F. X., GUNTHER, T. A. and TAY, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *Internat. Econom. Rev.* **39** 863–883.

DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econom. Statist.* **13** 253–263.

DIEBOLD, F. X., SCHUERMANN, T. and STROUGHAIR, J. D. (2000). Pitfalls and opportunities in the use of extreme value theory in risk management. *J. Risk Finance* **1** 30–35.

EFRON, B. (1991). Regression percentiles using asymmetric squared error loss. *Statist. Sinica* **1** 93–125. MR1101317

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. *Monographs on Statistics and Applied Probability* **57**. Chapman & Hall, New York. MR1270903

EHM, W., GNEITING, T., JORDAN, A. and KRÜGER, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 505–562. MR3506792

EMBRECHTS, P., KLÜPPELBERG, C. and MIKOSCH, T. (1997). *Modelling Extremal Events for Insurance and Finance*. *Applications of Mathematics* (*New York*) **33**. Springer, Berlin. MR1458613

EMMER, S., KRATZ, M. and TASCHE, D. (2015). What is the best risk measure in practice? A comparison of standard measures. *J. Risk* **18** 31–60.

ENGLE, R. F. and MANGANELLI, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *J. Bus. Econom. Statist.* **22** 367–381. MR2091566

FERNÁNDEZ, C. and STEEL, M. F. J. (1998). On Bayesian modeling of fat tails and skewness. *J. Amer. Statist. Assoc.* **93** 359–371. MR1614601

FISSLER, T. and ZIEGEL, J. F. (2016). Higher order elicitability and Osband's principle. *Ann. Statist.* **44** 1680–1707. MR3519937

FISSLER, T., ZIEGEL, J. F. and GNEITING, T. (2016). Expected shortfall is jointly elicitable with value at risk—implications for backtesting. *Risk Mag*. January 58–61.

FÖLLMER, H. and SCHIED, A. (2002). Convex measures of risk and trading constraints. *Finance Stoch.* **6** 429–447. MR1932379

FRONGILLO, R. and KASH, I. (2015). Vector-valued property elicitation. In *Proceedings of the* 28*th Conference on Learning Theory* (S. Kale, P. Grünwald and E. Hazan, eds.). *JMLR Workshop and Conference Proceedings* **40**.

GIACOMINI, R. and WHITE, H. (2006). Tests of conditional predictive ability. *Econometrica* **74** 1545–1578. MR2268409

GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. MR2847988

GNEITING, T. and RANJAN, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econom. Statist.* **29** 411–422. MR2848512

HOLZMANN, H. and EULERT, M. (2014). The role of the information set for forecasting—with applications to risk management. *Ann. Appl. Stat.* **8** 595–621. MR3192004

HOMMEL, G. (1983). Tests of the overall hypothesis for arbitrary dependence structures. *Biom. J.* **25** 423–430. MR0735888

KOENKER, R. (2005). *Quantile Regression*. *Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657

KOU, S. and PENG, X. (2016). On the measurement of economic tail risk. *Oper. Res.* **64** 1056–1072. MR3558435

KUAN, C.-M., YEH, J.-H. and HSU, Y.-C. (2009). Assessing value at risk with CARE, the conditional autoregressive expectile models. *J. Econometrics* **150** 261–270. MR2535521

KUESTER, K., MITTNIK, S. and PAOLELLA, M. S. (2006). Value-at-risk prediction: A comparison of alternative strategies. *J. Financ. Econom.* **4** 53–89.

LAMBERT, N. (2013). Elicitation and evaluation of statistical functionals. Preprint. Available at https://web.stanford.edu/~nlambert/papers/elicitation_july2013.pdf.

LAMBERT, N., PENNOCK, D. M. and SHOHAM, Y. (2008). Eliciting properties of probability distributions. In *Proceedings of the* 9*th ACM Conference on Electronic Commerce* 129–138. Chicago, IL. Extended abstract.

MCNEIL, A. J. and FREY, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *J. Empir. Finance* **7** 271–300.

MCNEIL, A. J., FREY, R. and EMBRECHTS, P. (2005). *Quantitative Risk Management*: *Concepts*, *Techniques and Tools*. *Princeton Series in Finance*. Princeton Univ. Press, Princeton, NJ. MR2175089

NAU, R. F. (1985). Should scoring rules be "effective"? *Manage. Sci.* **31** 527–535.

NEWEY, W. K. and POWELL, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica* **55** 819–847. MR0906565

NOLDE, N. and ZIEGEL, J. F. (2017). Supplement to "Elicitability and backtesting: Perspectives for banking regulation". DOI:10.1214/17-AOAS1041SUPP.

OSBAND, K. H. (1985). Providing incentives for better cost forecasting. Ph.D. thesis, Univ. California, Berkeley.

PATTON, A. J. (2006). Volatility forecast comparison using imperfect volatility proxies. Research Paper 175, Quantitative Finance Research Centre, Univ. Technology, Sydney.

PATTON, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *J. Econometrics* **160** 246–256. MR2745881

PATTON, A. J. (2014). Evaluating and comparing possibly misspecified forecasts. Working paper.

PATTON, A. J. and SHEPPARD, K. (2009). Evaluating volatility and correlation forecasts. In *Handbook of Financial Time Series* (T. Mikosch, J.-P. Kreiss, R. A. Davis and T. G. Andersen, eds.) 801–838. Springer, Berlin.

R CORE TEAM (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at http://www.R-project.org/.

ROSENBLATT, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Stat.* **23** 470–472. MR0049525

SAERENS, M. (2000). Building cost functions minimizing to some summary statistics. *IEEE Trans. Neural Netw.* **11** 1263–1271.

STEINWART, I., PASIN, C., WILLIAMSON, R. and ZHANG, S. (2014). Elicitation and identification of properties. *J. Mach. Learn. Res. Workshop Conf. Proc.* **35** 1–45.

STRÄHL, C. and ZIEGEL, J. (2017). Cross-calibration of probabilistic forecasts. *Electron. J. Stat.* **11** 608–639. MR3619318

THOMSON, W. (1979). Eliciting production possibilities from a well-informed manager. *J. Econom. Theory* **20** 360–380. MR0540822

TSYPLAKOV, A. (2014). Theoretical guidelines for a partially informed forecast examiner. MPRA Paper, 55017. Available at http://mpra.ub.uni-muenchen.de/55017.

WANG, R. and ZIEGEL, J. F. (2015). Elicitable distortion risk measures: A concise proof. *Statist. Probab. Lett.* **100** 172–175. MR3324090

WEBER, S. (2006). Distribution-invariant risk measures, information, and dynamic consistency. *Math. Finance* **16** 419–441. MR2212272

ZIEGEL, J. F. (2016). Coherence and elicitability. *Math. Finance* **26** 901–918. MR3551510

DEPARTMENT OF STATISTICS
UNIVERSITY OF BRITISH COLUMBIA
VANCOUVER, BRITISH COLUMBIA V6T 1Z4
CANADA
E-MAIL: natalia@stat.ubc.ca

INSTITUTE OF MATHEMATICAL STATISTICS
AND ACTUARIAL SCIENCE
UNIVERSITY OF BERN
CH-3012 BERN
SWITZERLAND
E-MAIL: johanna.ziegel@stat.unibe.ch