The Annals of Applied Statistics
2023, Vol. 17, No. 3, 2236–2255
https://doi.org/10.1214/22-AOAS1716

USING PROXIES TO IMPROVE FORECAST EVALUATION

BY HAJO HOLZMANN[1,a] AND BERNHARD KLAR[2,b]

[1]*Fachbereich Mathematik und Informatik, Philipps-Universität Marburg,* [a]*holzmann@mathematik.uni-marburg.de*
[2]*Institut für Stochastik, Karlsruher Institut für Technologie (KIT),* [b]*bernhard.klar@kit.edu*

Comparative evaluation of forecasts of statistical functionals relies on comparing averaged losses of competing forecasts after the realization of the quantity Y, on which the functional is based, has been observed. Motivated by high-frequency finance, in this paper we investigate how proxies $\tilde{Y}$ for Y—say volatility proxies—which are observed together with Y can be utilized to improve forecast comparisons. We extend previous results on robustness of loss functions for the mean to general moments and ratios of moments, and show in terms of the variance of differences of losses that using proxies will increase the power in comparative forecast tests. These results apply both to testing conditional as well as unconditional dominance. Finally, we numerically illustrate the theoretical results, both for simulated high-frequency data as well as for high-frequency log returns of several cryptocurrencies.

1. Introduction. Comparative evaluation of forecasts of statistical functionals is a standard issue in the realm of forecasting (Gneiting (2011)). It relies on comparing expected or averaged losses of competing forecasts, after the realization of the quantity Y, on which the functional is based, has been observed. The aim of this paper is to investigate how proxies $\tilde{Y}$ for Y, which are observed together with Y, can be utilized to improve forecast comparisons in the sense that they result in the same ordering but bring an increase in the power of comparative forecast tests.

The motivation comes mainly from high-frequency finance, where high-frequency data are routinely used to *generate* forecasts—say of volatilities—also over moderate time horizons such as daily volatilities (Corsi (2009)). Our investigation shows how high-frequency data can be used to obtain *sharper forecast evaluation*. In comparative forecast evaluation, this would mean that when comparing two forecasts of daily volatilities in terms of expected values of loss functions, the better forecast can be determined with higher power when using these high-frequency data in the process of forecast evaluation.

Our main point of departure was Patton (2011), who showed that using various noisy volatility proxies, for example, based on high frequency, is *valid* in comparative forecast evaluation, that is, preserves the order of the expected losses. Hansen and Lunde (2006) have similar results, while Laurent, Rombouts and Violante (2013) provide a multivariate generalization of the characterization in Patton (2011) and Koopman, Jungbacker and Hol (2005) illustrate the use of realized measures for forecast comparisons on various observed high frequency data sets. We are interested in the comparison of different. possibly misspecified forecasts, in which situation Patton (2020) shows that the ranking of the forecasts may depend on the loss function.

A very recent, related contribution is by Hoga and Dimitriadis (2022), who focus on predicting the mean, and illustrate their methods for GDP forecasts. Our contributions and their relation to the literature can be summarized as follows:

(i) We extend the analysis from Patton (2011) and Hansen and Lunde (2006) about the validity when using various proxies from volatilities, that is, second moments, to general

Received December 2021; revised October 2022.
Key words and phrases. Cryptocurrencies, DM test, forecast comparison, volatility proxies.

moments and beyond to ratios of moments. We use a concept corresponding to the notion of exact robustness from Hoga and Dimitriadis (2022), who also assume that the proxy enters the loss difference in the same way as the original observation, and in this setting show that only the mean allows for exactly robust loss functions.

(ii) We formally show in terms of the variance of differences of losses that using proxies will increase the power in comparative forecast testing by decreasing the variance of loss differences, both when testing conditional as well as unconditional dominance; see Nolde and Ziegel (2017)) for these notions. Hoga and Dimitriadis (2022) have similar results for the mean and focus on conditional dominance testing.

(iii) Finally, we illustrate the theoretical results for high-frequency data, both simulated as well as related to three cryptocurrencies, using the three-zone approach from Fissler, Ziegel and Gneiting (2016) and Nolde and Ziegel (2017). We show that the choice of the proxies, as well as the choice of the loss function, has a pronounced effect on the comparative evaluation of forecasts; using high-frequency data and the QLIKE loss substantially improves the forecast evaluation.

The paper is structured as follows. In Section 2, we start with a motivating example; we recall strictly consistent loss functions for statistical functionals, and introduce a dynamic framework for forecast evaluation. Section 3 investigates the use of proxies to improve evaluations of forecasts of moments, and in Section 4 this is further discussed and extended to ratios of moments. Section 6 summarizes the results of a simulation study, where we consider comparing forecasts for second, third, and fourth moments for GARCH-type time series. In Section 7, we provide an illustration of our methods to predicting the volatility of log-returns of three cryptocurrencies, while a supplement contains additional numerical results.

2. Motivation and basic concepts.

2.1. *Motivating examples*. Let us first illustrate the use of different proxies and loss functions in Diebold–Mariano (DM) tests for equal forecast performance. Consider the following stylized scenario, where the aim is to distinguish between two competing forecasts. Observations correspond to logarithmic returns, and we assume that the true data generating process is a simple GARCH(1,1) model. The total length of the time series is $T = 1500$, and we consider rolling one-step-ahead forecasts of the conditional variance using a moving time window, with window length T/3, resulting in $n = 1000$ forecasts.

There are four forecasters. The first one is lucky to use a GARCH(1,1) model for making predictions, forecasters 2, 3, and 4 use ARCH(1), ARCH(2), and ARCH(7) models, respectively. Clearly, we expect that the predictions from forecaster 1 outperform, in some sense, the other predictions. Moreover, we would expect that the ARCH(7) model beats the ARCH(1) and ARCH(2) models since the former should be a better approximation to a GARCH(1,1) process. Let the forecasts of the conditional variance of logarithmic returns r_t for any pair of forecasters be denoted by $x_{1,t}$ and $x_{2,t}$. Then our interest focuses on the null hypotheses,

$$H_0: \quad \text{Forecast } x_{1,t} \text{ predicts at least as well as forecast } x_{2,t},$$

and if H_0 is rejected, then $x_{2,t}$ is worse than $x_{1,t}$. To decide for or against H_0, we use a DM test based on the loss differences

$$\Delta_n \bar{L} = 1/n \sum_{t=1}^{n} L(x_{1,t}, y_{t+1}) - L(x_{2,t}, y_{t+1}),$$

where $L(x, y)$ is some loss function, and y_{t+1} materializes at day $t + 1$. Under suitable conditions, the studentized test statistic S has a limiting standard normal distribution, and H_0 is rejected for large values of S.

To evaluate the forecasts, the mean squared error loss $L(x, y) = (x - y)^2$, is used, together with the squared returns $\tilde{y}_t = r_t^2$ as an unbiased proxy for the true conditional variance of logarithmic returns. The first line of the following table shows the values of the test statistic S if the different predictions are compared to the GARCH(1,1) model. Even if the values are positive (hence, slightly favor the GARCH model), they are nowhere near statistically significant. The comparison of the ARCH(7) model with the other two ARCH models even results in values close to zero:

$x_{2,t}$ \ $x_{1,t}$	GARCH(1,1)	ARCH(1)	ARCH(2)	ARCH(7)
GARCH(1,1)	–	0.788	1.010	0.984
ARCH(7)	−0.984	−0.193	0.152	–

Next, assume that, besides the daily returns, also 5-min returns are available for predicting the next day's volatility. Thus, the squared returns are replaced by the realized variances $\tilde{y}_t = \sum_{i=1}^{m} r_{t,i}^2$, where $r_{t,i}$ are the intraday log returns. The outcomes of the DM tests for equal predictive performance, now using the realized variances as proxies, are as follows:

$x_{2,t}$ \ $x_{1,t}$	GARCH(1,1)	ARCH(1)	ARCH(2)	ARCH(7)
GARCH(1,1)	–	3.976	3.924	3.506
ARCH(7)	−3.506	2.474	2.352	–

In comparison with the first table, the values in the first line are much larger, being statistically significant even on the 0.01-level, and indicating the dominance of the prediction under the GARCH(1,1) model. The comparison of the ARCH(7) model with the other two ARCH models favors the ARCH(7) model, at least on the 0.05-level.

Finally, the evaluator decides to replace the MSE with the QLIKE loss function $\tilde{L}(x, y) = y/x - \log(y/x) - 1$, and gets the following results of the corresponding DM tests:

$x_{2,t}$ \ $x_{1,t}$	GARCH(1,1)	ARCH(1)	ARCH(2)	ARCH(7)
GARCH(1,1)	–	5.589	5.523	4.386
ARCH(7)	−4.386	3.642	3.492	–

Now, all entries are even larger in absolute terms, and the ARCH(7) model dominates the competing ARCH models even on the 0.01-level.

Clearly, these results are based on a specific realization of the time series. However, a closer look at this example in Section 6 reveals that this behavior is rather typical.

As a real data example, we consider log returns of the cryptocurrency Bitcoin (BTC). We use hourly observations from May 16, 2018, to October 27, 2021, with sample size 30264, which corresponds to 1261 days. All prices are closing values in U.S. dollars. Returns are estimated by taking logarithmic differences. Figure 1 shows the results of DM tests for the log returns for different competing models, where the results are depicted using the three-zone approach of Fissler, Ziegel and Gneiting (2016). A result marked in green indicates that the model in the column outperforms the model in the row, a red mark indicates inferiority. The marking is yellow if there is no significant difference between the two models. Light green, green, and dark green indicates significance at level 0.1, 0.05, and 0.01, respectively. For example, in the upper left panel of Figure 1, the CGARCH model outperforms the ARCH(1) model, but none of the other models. Whereas there are only slight differences between the use of squared returns compared to high frequency data as proxies in case of MSE, the power of the DM tests is considerably higher by using the high frequency proxy compared to squared returns with QLIKE loss. Moreover, there is an increase in power using the QLIKE loss function compared to MSE. For more details, see Sections 5 and 7.

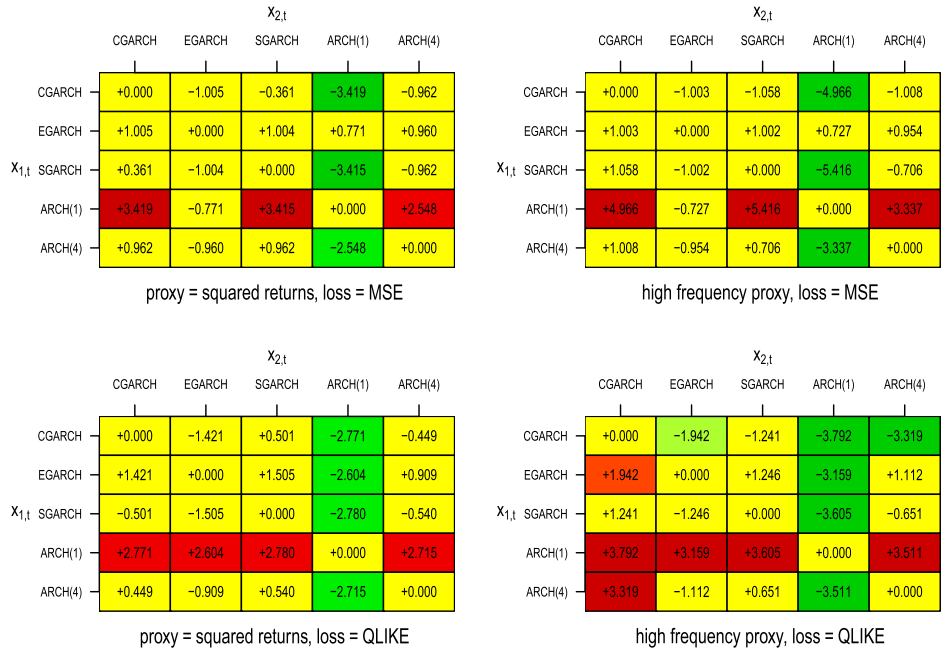

FIG. 1. *Results of DM tests for BTC log returns from May* 16, 2018, *to October* 27, 2021, *left*: *squared returns*, *right*: *high frequency proxy.*

2.2. *Loss functions and statistical functionals.* We start by recalling the concept of strictly consistent loss (or scoring) functions; see Gneiting (2011). Let Θ be a class of distribution functions on a closed subset $D \subset \mathbb{R}$, which we identify with their associated probability distributions, and let $T : \Theta \to \mathbb{R}$ be a (one-dimensional) *statistical functional (or parameter)*.

A *loss function (also scoring function)* is a measurable map $\mathrm{L} : \mathbb{R} \times D \to [0, \infty)$. It is interpreted as the loss if forecast x is issued and y materializes. L is consistent for the functional T relative to the class Θ, if

$$\text{for all } x \in \mathbb{R},\ F \in \Theta : \quad \mathbb{E}_F\big[\mathrm{L}\big(T(F), Y\big)\big] \le \mathbb{E}_F\big[\mathrm{L}(x, Y)\big], \tag{1}$$

and $\mathbb{E}_F$ indicates that expectation is taken under the distribution F for Y, and we assume that the relevant expected values are finite. Thus, the true functional $T(F)$ minimizes the expected loss under F. If, in addition,

$$\mathbb{E}_F\big[\mathrm{L}\big(T(F), Y\big)\big] = \mathbb{E}_F\big[\mathrm{L}(x, Y)\big] \quad \text{implies that } x = T(F),$$

then L is *strictly consistent* for T. The functional T is called *elicitable* relative to the class Θ if it admits a strictly consistent loss function. For several functionals such as moments, quantiles, and expectiles, Gneiting (2011) characterizes all strictly-consistent loss functions under some smoothness and normalization conditions; see also Steinwart et al. (2014).

When comparing two forecasts $x, x' \in \mathbb{R}$ for a given $F \in \Theta$, and hence parameter $T(F)$, we say that x *dominates* x' *under* F *for the loss* L if the *difference of expected losses* is negative:

$$\mathbb{E}_F\big[\mathrm{L}(x, Y)\big] - \mathbb{E}_F\big[\mathrm{L}\big(x', Y\big)\big] < 0. \tag{2}$$

From (1), for a strictly-consistent loss function, the true parameter $T(F)$ dominates any other forecast.

2.3. *Dynamic forcasting and comparative forecast evaluation.* Now let us consider a forecasting situation. Forecasts are issued on the basis of certain information. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, let $\mathcal{F}_t \subset \mathcal{A}$ be a sub-σ-algebra of $\mathcal{A}$, the information set at time t, on the basis of which the forecast is issued. In finance, $\mathcal{F}_t$ can include returns (including high-frequency) up to time t as well other covariates observed up to time t.

The aim is to predict the functional T, say the mean or the volatility, of the random variable $Y_{t+1} : \Omega \to \mathbb{R}$, which will be observed at time $t+1$ (say one day ahead). For example, this may be the return from t to $t+1$ over one day. More precisely, if $F_{Y_{t+1}|\mathcal{F}_t}(\omega, \cdot)$ denotes the conditional distribution of Y_{t+1} given $\mathcal{F}_t$, then the parameter of interest is

$$T(Y_{t+1}|\mathcal{F}_t)(\omega) := T\big(F_{Y_{t+1}|\mathcal{F}_t}(\omega, \cdot)\big)$$

We note that

- the forecast is based on the full information up to time t. Thus, even if Y_{t+1} is a return over one day, for the forecast we use, for example, high-frequency data up to time t if these are included in $\mathcal{F}_t$,
- compared to such additional data, the observation Y_{t+1} is of particular relevance since the parameter $T(Y_{t+1}|\mathcal{F}_t)$ is defined via its conditional distribution $F_{Y_{t+1}|\mathcal{F}_t}(\omega, \cdot)$ given $\mathcal{F}_t$.

Thus, to generate the forecast, even if the time horizon for the forecast is, say one day, it is standard to use high-frequency information contained in $\mathcal{F}_t$ up to time t. Now, the issue in this paper is how to use additional information contained in $\mathcal{F}_{t+1}$, available at time $t+1$, to improve comparative forecast evaluation.

First, let us recall the setting for comparative forecast evaluation based on Y_{t+1}. A forecast at time t is—in great generality—an $\mathcal{F}_t$-measurable random variable Z_t. Now, if L is a strictly consistent loss function for T, then compared to the true forecast $T(Y_{t+1}|\mathcal{F}_t)$, we have the following results (Holzmann and Eulert (2014)):

$$\mathbb{E}\big[\mathrm{L}\big(T(Y_{t+1}|\mathcal{F}_t), Y_{t+1}\big)|\mathcal{F}_t\big](\omega) \le \mathbb{E}\big[\mathrm{L}(Z_t, Y_{t+1})|\mathcal{F}_t\big](\omega) \quad \text{for } \mathbb{P}\text{-a.e. } \omega \in \Omega, \tag{3}$$

that is, the conditional dominance of the true forecast $T(Y_{t+1}|\mathcal{F}_t)$ over any other generic forecast Z_t, and

$$\mathbb{E}\big[\mathrm{L}\big(T(Y_{t+1}|\mathcal{F}_t), Y_{t+1}\big)\big] \le \mathbb{E}\big[\mathrm{L}(Z_t, Y_{t+1})\big], \tag{4}$$

the unconditional dominance of $T(Y_{t+1}|\mathcal{F}_t)$ over Z_t. When comparing with the true forecast $T(Y_{t+1}|\mathcal{F}_t)$ (but not in general), these two concepts coincide: We have equality in (3) or (4) if and only if $T(Y_{t+1}|\mathcal{F}_t) = Z_t$ $\mathbb{P}$-almost surely. Note that Y_{t+1} is used in the comparisons (3) and (4) by default.

We shall generally compare two potentially misspecified forecasts, that is, general $\mathcal{F}_t$-measurable random variables Z_t and Z_t' none of which needs to coincide with $T(Y_{t+1}|\mathcal{F}_t)$. Then, by definition, Z_t' *conditionally dominates* Z_t for the loss function L if

$$\mathbb{E}\big[\mathrm{L}\big(Z_t', Y_{t+1}\big)|\mathcal{F}_t\big](\omega) \le \mathbb{E}\big[\mathrm{L}(Z_t, Y_{t+1})|\mathcal{F}_t\big](\omega) \quad \text{for } \mathbb{P}\text{-a.e. } \omega \in \Omega, \tag{5}$$

with strict inequality on a set of positive probability, while Z_t' *unconditionally dominates* Z_t for the loss function L if

$$\mathbb{E}\big[\mathrm{L}\big(Z_t', Y_{t+1}\big)\big] < \mathbb{E}\big[\mathrm{L}(Z_t, Y_{t+1})\big]. \tag{6}$$

In this more general setting, conditional dominance still implies unconditional dominance, but the converse is not true in general.

Now, we consider how additional information contained in $\mathcal{F}_{t+1}$ (apart from Y_{t+1}) may be used for *forecast evaluation*. In the context of high frequency financial data, apart from using the high-frequency data to *generate* forecasts over daily time horizons, we shall investigate these high-frequency data to obtain a *sharper forecast evaluation*.

3. Proxies when comparing forecasts of moments. Suppose that $D = I$ is an interval, $h : I \to \mathbb{R}$ is a measurable function such that $\mathbb{E}_F[|h(Y)|] < \infty$ for all $F \in \Theta$. Then a classical result by Savage (1971) (see also Gneiting (2011)), characterizes the strictly consistent scoring functions for $T(F) = \mathbb{E}_F[h(Y)]$ in the form

$$\mathrm{L}(x, y) = \phi(y) - \phi(x) - \phi'(x)\big(h(y) - x\big), \quad y, x \in I, \tag{7}$$

where $\phi : I \to \mathbb{R}$ is a strictly convex function with subgradient ϕ' for which $\mathbb{E}_F[|\phi(X)|] < \infty$ for all $F \in \Theta$, and ϕ' denotes the derivative of ϕ.

The functional $T(F) = \mathbb{E}_F[h(Y)]$ is called the *h-moment* of Y, or *generalized moment* or simply a moment of Y. The classical moments arise for $h(x) = x^n$ for integers n. In our applications, we shall focus on the cases $n = 2$, 3, and $n = 4$.

First, we formulate the following lemma in the static framework.

LEMMA 1. *Consider* (7) *and forecasts* $x_1, x_2 \in \mathbb{R}$.

(i) *The loss difference*,

$$\begin{aligned}\mathrm{L}(x_1, y) - \mathrm{L}(x_2, y) &= \phi(x_2)(1 - x_2) - \phi(x_1)(1 - x_1) + \big(\phi'(x_2) - \phi'(x_1)\big)h(y) \\ &=: \mathrm{L}_{\mathrm{Diff}}\big(x_1, x_2, h(y)\big)\end{aligned} \tag{8}$$

depends on y *only through* $h(y)$.

(ii) *If* $F \in \Theta$ *and* $\tilde{Y}$ *is a random variable* (*with a given distribution*) *such that* $\mathbb{E}_F[h(Y)] = \mathbb{E}[\tilde{Y}]$ (*the moment is the same*), *then*

$$\mathbb{E}_F\big[\mathrm{L}_{\mathrm{Diff}}\big(x_1, x_2, h(Y)\big)\big] = \mathbb{E}\big[\mathrm{L}_{\mathrm{Diff}}(x_1, x_2, \tilde{Y})\big]. \tag{9}$$

(iii) *We have that*

$$\mathrm{Var}_F\big(\mathrm{L}_{\mathrm{Diff}}\big(x_1, x_2, h(Y)\big)\big) = \big(\phi'(x_2) - \phi'(x_1)\big)^2 \mathrm{Var}_F\big(h(Y)\big). \tag{10}$$

Consequently, if in addition to (*ii*) *it holds that* $\mathrm{Var}(\tilde{Y}) \le \mathrm{Var}_F[h(Y)]$, *then we have that*

$$\mathrm{Var}\big(\mathrm{L}_{\mathrm{Diff}}(x_1, x_2, \tilde{Y})\big) \le \mathrm{Var}_F\big(\mathrm{L}_{\mathrm{Diff}}\big(x_1, x_2, h(Y)\big)\big). \tag{11}$$

Here, $\tilde{Y}$ plays the role of the proxy that shall be used to improve forecast evaluation. Part (ii) shows that using $\tilde{Y}$ instead of $h(Y)$ is *valid* if $\mathbb{E}_F[h(Y)] = \mathbb{E}[\tilde{Y}]$ in the sense that dominance relations of forecasts are preserved when using $\tilde{Y}$, while (11) shows that evaluation of score differences is actually sharper based on $\tilde{Y}$ instead of $h(Y)$ if $\mathrm{Var}(\tilde{Y}) \le \mathrm{Var}_F[h(Y)]$.

Hoga and Dimitriadis (2022) call the equality of loss differences in (9) exact robustness. When assuming that the proxy $\tilde{Y}$ enters the loss difference in the same fashion as Y, they show that exact robustness can only hold for strictly consistent scoring functions of the mean. In our more flexible approach, we cover general moments and also ratios of moments; see below.

PROOF. Part (i) is easily checked by inserting the loss function (7).

Concerning part (ii), inserting $\mathrm{L}_{\mathrm{Diff}}$ from (8), we get by assumption

$$\begin{aligned}&\mathbb{E}_F\big[\mathrm{L}_{\mathrm{Diff}}\big(x_1, x_2, h(Y)\big)\big] - \mathbb{E}\big[\mathrm{L}_{\mathrm{Diff}}(x_1, x_2, \tilde{Y})\big] \\ &\quad = \big(\phi'(x_2) - \phi'(x_1)\big)\big(\mathbb{E}_F\big[h(Y)\big] - \mathbb{E}[\tilde{Y}]\big) = 0\end{aligned}$$

Part (iii) follows similarly easily. □

Now, let us turn to the dynamic setting described in Section 2.3.

THEOREM 2 (Forecast dominance testing). *Consider forecasting the conditional moment* $\mathbb{E}[h(Y_{t+1})|\mathcal{F}_t]$, *and suppose that* $\tilde{Y}_{t+1}$ *is* $\mathcal{F}_{t+1}$*-measurable with*

$$\mathbb{E}[\tilde{Y}_{t+1}|\mathcal{F}_t] = \mathbb{E}\big[h(Y_{t+1})|\mathcal{F}_t\big] \quad a.s. \tag{12}$$

(i) *For the loss difference* (8), *for any two forecasts* Z_t *and* Z_t' ($\mathcal{F}_t$*-measurable random variables*),

$$\mathbb{E}\big[\mathrm{L}_{\mathrm{Diff}}(Z_t, Z_t', h(Y_{t+1}))|\mathcal{F}_t\big] = \mathbb{E}\big[\mathrm{L}_{\mathrm{Diff}}(Z_t, Z_t', \tilde{Y}_{t+1})|\mathcal{F}_t\big], \tag{13}$$

and hence in particular

$$\mathbb{E}\big[\mathrm{L}_{\mathrm{Diff}}(Z_t, Z_t', h(Y_{t+1}))\big] = \mathbb{E}\big[\mathrm{L}_{\mathrm{Diff}}(Z_t, Z_t', \tilde{Y}_{t+1})\big]. \tag{14}$$

Thus, both conditional as well as unconditional dominance are preserved when using $\tilde{Y}_{t+1}$ *instead of* $h(Y_{t+1})$ *in the forecast comparison.*

(ii) *If in addition to* (12), *we have that*

$$\mathrm{Var}(\tilde{Y}_{t+1}|\mathcal{F}_t) \le \mathrm{Var}\big(h(Y_{t+1})|\mathcal{F}_t\big),$$

then

$$\mathrm{Var}\big(\mathrm{L}_{\mathrm{Diff}}(Z_t, Z_t', \tilde{Y}_{t+1})|\mathcal{F}_t\big) \le \mathrm{Var}\big(\mathrm{L}_{\mathrm{Diff}}(Z_t, Z_t', h(Y_{t+1}))|\mathcal{F}_t\big) \tag{15}$$

as well as

$$\mathrm{Var}\big(\mathrm{L}_{\mathrm{Diff}}(Z_t, Z_t', \tilde{Y}_{t+1})\big) \le \mathrm{Var}\big(\mathrm{L}_{\mathrm{Diff}}(Z_t, Z_t', h(Y_{t+1}))\big). \tag{16}$$

The second part of the theorem shows that a variance reduction is achieved both for testing conditional as well as unconditional dominance.

PROOF. (i): Equation (13) is (9) in Lemma 1, conditional on $\mathcal{F}_t$, while (14) follows from (13) by taking expected values.

(ii): Inequality (15) is (11) in Lemma 1, (iii), conditional on $\mathcal{F}_t$. As for (16), we have that

$$\mathrm{Var}\big(\mathrm{L}_{\mathrm{Diff}}(Z_t, Z_t', \tilde{Y}_{t+1})\big) = \mathbb{E}\big[\mathrm{Var}\big(\mathrm{L}_{\mathrm{Diff}}(Z_t, Z_t', \tilde{Y}_{t+1})|\mathcal{F}_t\big)\big] + \mathrm{Var}\big(\mathbb{E}\big[\mathrm{L}_{\mathrm{Diff}}(Z_t, Z_t', \tilde{Y}_{t+1})|\mathcal{F}_t\big]\big)$$

Since by (13),

$$\mathrm{Var}\big(\mathbb{E}\big[\mathrm{L}_{\mathrm{Diff}}(Z_t, Z_t', \tilde{Y}_{t+1})|\mathcal{F}_t\big]\big) = \mathrm{Var}\big(\mathbb{E}\big[\mathrm{L}_{\mathrm{Diff}}(Z_t, Z_t', h(Y_{t+1}))|\mathcal{F}_t\big]\big)$$

the conclusion follows since

$$\begin{aligned}\mathbb{E}\big[\mathrm{Var}\big(\mathrm{L}_{\mathrm{Diff}}(Z_t, Z_t', \tilde{Y}_{t+1})|\mathcal{F}_t\big)\big] &= \mathbb{E}\big[\big(\phi'(Z_t') - \phi'(Z_t)\big)^2 \mathrm{Var}(\tilde{Y}_{t+1}|\mathcal{F}_t)\big] \\ &\le \mathbb{E}\big[\big(\phi'(Z_t') - \phi'(Z_t)\big)^2 \mathrm{Var}\big(h(Y_{t+1})|\mathcal{F}_t\big)\big] \\ &= \mathbb{E}\big[\mathrm{Var}\big(\mathrm{L}_{\mathrm{Diff}}(Z_t, Z_t', h(Y_{t+1}))|\mathcal{F}_t\big)\big]. \qquad \square\end{aligned}$$

4. Ratios of moments and further parameters. Suppose that $D = I$ is an interval, $h : I \to \mathbb{R}$ and $s : I \to (0, \infty)$ are measurable functions such that $\mathbb{E}_F[|h(Y)|] < \infty$, $\mathbb{E}_F[s(Y)] < \infty$ for all $F \in \Theta$. The target parameter is

$$T(F) = \frac{\mathbb{E}_F[h(Y)]}{\mathbb{E}_F[s(Y)]}.$$

Gneiting (2011) shows that strictly consistent loss functions for $T(F)$ are of the form

$$\mathrm{L}(x, y) = s(y)\big(\phi(y) - \phi(x)\big) - \phi'(x)\big(h(y) - xs(y)\big) - \phi'(y)\big(h(y) - ys(y)\big) \tag{17}$$

($y, x \in I$), where it is additionally assumed that

$$\mathbb{E}_F\big[|h(Y)||\phi'(Y)|\big] < \infty, \mathbb{E}_F\big[|s(Y)||\phi(Y)|\big] < \infty, \mathbb{E}_F\big[|Y||s(Y)||\phi(Y)|\big] < \infty, \quad F \in \Theta.$$

LEMMA 3. *Consider* (17) *and forecasts* $x_1, x_2 \in \mathbb{R}$.

(i) *The loss difference*,

$$\begin{aligned} &\mathrm{L}(x_1, y) - \mathrm{L}(x_2, y) \\ &\quad = \big(\phi'(x_2) - \phi'(x_1)\big)h(y) + \big(x_1\phi'(x_1) - \phi(x_1) - x_2\phi'(x_2) + \phi(x_2)\big)s(y) \\ &\quad =: \mathrm{L}_{\mathrm{Diff}}\big(x_1, x_2, h(y), s(y)\big) \end{aligned} \tag{18}$$

depends on y *only through* $h(y)$ *and* $s(y)$.

(ii) *If* $F \in \Theta$, *and* $\tilde{Y}_1, \tilde{Y}_2$ *are random variables* (*with given distributions*) *such that* $\mathbb{E}_F[h(Y)] = \mathbb{E}[\tilde{Y}_1]$ *and* $\mathbb{E}_F[s(Y)] = \mathbb{E}[\tilde{Y}_2]$ (*the moment is the same*), *then*

$$\mathbb{E}_F\big[\mathrm{L}_{\mathrm{Diff}}\big(x_1, x_2, h(Y), s(Y)\big)\big] = \mathbb{E}\big[\mathrm{L}_{\mathrm{Diff}}(x_1, x_2, \tilde{Y}_1, \tilde{Y}_2)\big]. \tag{19}$$

(iii) *If in addition to* (*ii*) *we have that* $\mathrm{Var}(a\tilde{Y}_1 + b\tilde{Y}_2) \le \mathrm{Var}_F(ah(Y) + bs(Y))$ *for* $a, b \in \mathbb{R}$ *we have that*

$$\mathrm{Var}\big(\mathrm{L}_{\mathrm{Diff}}(x_1, x_2, \tilde{Y}_1, \tilde{Y}_2)\big) \le \mathrm{Var}_F\big(\mathrm{L}_{\mathrm{Diff}}\big(x_1, x_2, h(Y), s(Y)\big)\big). \tag{20}$$

PROOF. The form (18) of the loss difference follows directly from inserting (17). Then (19) and (20) follow immediately from the form of the loss difference. □

Note that the result of the lemma does not contradict Theorem 1 in Hoga and Dimitriadis (2022), since they only allow a single proxy $\hat{Y}_t$, which enters the loss function in the same way as Y_t, whereas in the above lemma we require proxies $\tilde{Y}_1$ and $\tilde{Y}_2$ for the two moments $h(Y)$ and $s(Y)$.

THEOREM 4 (Forecast dominance testing: Ratio of moments). *Consider forecasting the ratio of conditional moments* $\mathbb{E}[h(Y_{t+1})|\mathcal{F}_t]/\mathbb{E}[s(Y_{t+1})|\mathcal{F}_t]$, *and suppose that* $\tilde{Y}_{t+1}^{(j)}$ *are* $\mathcal{F}_{t+1}$*-measurable with*

$$\mathbb{E}\big[\tilde{Y}_{t+1}^{(1)}|\mathcal{F}_t\big] = \mathbb{E}\big[h(Y_{t+1})|\mathcal{F}_t\big], \qquad \mathbb{E}\big[\tilde{Y}_{t+1}^{(2)}|\mathcal{F}_t\big] = \mathbb{E}\big[s(Y_{t+1})|\mathcal{F}_t\big] \quad a.s. \tag{21}$$

(i) *For the loss difference* (18), *for any two forecasts* Z_t *and* Z_t' ($\mathcal{F}_t$*-measurable random variables*),

$$\mathbb{E}\big[\mathrm{L}_{\mathrm{Diff}}\big(Z_t, Z_t', h(Y_{t+1}), s(Y_{t+1})\big)|\mathcal{F}_t\big] = \mathbb{E}\big[\mathrm{L}_{\mathrm{Diff}}\big(Z_t, Z_t', \tilde{Y}_{t+1}^{(1)}, \tilde{Y}_{t+1}^{(2)}\big)|\mathcal{F}_t\big], \tag{22}$$

and hence in particular

$$\mathbb{E}\big[\mathrm{L}_{\mathrm{Diff}}\big(Z_t, Z_t', h(Y_{t+1}), s(Y_{t+1})\big)\big] = \mathbb{E}\big[\mathrm{L}_{\mathrm{Diff}}\big(Z_t, Z_t', \tilde{Y}_{t+1}^{(1)}, \tilde{Y}_{t+1}^{(2)}\big)\big]. \tag{23}$$

(ii) *If in addition to* (21) *we have that for all* $\mathcal{F}_t$*-measurable random variables* V, W, *we have that*

$$\begin{aligned} &V^2\,\mathrm{Var}\big(\tilde{Y}_{t+1}^{(1)}|\mathcal{F}_t\big) + W^2\,\mathrm{Var}\big(\tilde{Y}_{t+1}^{(2)}|\mathcal{F}_t\big) + 2VW\,\mathrm{Cov}\big(\tilde{Y}_{t+1}^{(1)}, \tilde{Y}_{t+1}^{(2)}|\mathcal{F}_t\big) \\ &\quad \le V^2\,\mathrm{Var}\big(h(Y_{t+1})|\mathcal{F}_t\big) + W^2\,\mathrm{Var}\big(s(Y_{t+1})|\mathcal{F}_t\big) + 2VW\,\mathrm{Cov}\big(h(Y_{t+1}), s(Y_{t+1})|\mathcal{F}_t\big), \end{aligned}$$

then

$$\mathrm{Var}\big(\mathrm{L}_{\mathrm{Diff}}\big(Z_t, Z_t', \tilde{Y}_{t+1}^{(1)}, \tilde{Y}_{t+1}^{(2)}\big)|\mathcal{F}_t\big) \le \mathrm{Var}\big(\mathrm{L}_{\mathrm{Diff}}\big(Z_t, Z_t', h(Y_{t+1}), s(Y_{t+1})\big)|\mathcal{F}_t\big) \tag{24}$$

as well as

$$\mathrm{Var}\big(\mathrm{L}_{\mathrm{Diff}}\big(Z_t, Z_t', \tilde{Y}_{t+1}^{(1)}, \tilde{Y}_{t+1}^{(2)}\big)\big) \le \mathrm{Var}\big(\mathrm{L}_{\mathrm{Diff}}\big(Z_t, Z_t', h(Y_{t+1}), s(Y_{t+1})\big)\big) \tag{25}$$

The proof is immediate from Lemma 3, and the final inequality (25) follows as (16) in Theorem 2. The condition for a potential variance reduction in Theorem 4(ii) is more restrictive than that from Theorem 2(ii), and, apart from relating the variances of $s(Y)$ and $h(Y)$ to those of the proxies, also involves conditional covariances.

Theorem 4 does not apply to measures such as skewness and kurtosis, which even for centered distributions are known not to allow for strictly consistent scoring functions. However, the revelation principle, Theorem 4 in Gneiting (2011), and the elicitability (existence of a strictly consistent scoring function), and hence joint elicitability of moments implies that for centered distributions, these measures are elicitable when considered together with the second moment. Roughly speaking, for the skewness this involves the two-dimensional parameter consisting of third and second moment, and for the kurtosis consisting of fourth and second moment. The analysis of the corresponding loss differences is then similar to that in Theorem 4.

5. Diebold–Mariano testing. We briefly summarize the DM test (Diebold and Mariano (1995)) for forecast dominance, where we shall focus on unconditional dominance. For a discussion of conditional dominance testing, together with asymptotic theory and local power analysis see Hoga and Dimitriadis (2022).

As proposed in Fissler, Ziegel and Gneiting (2016) and Nolde and Ziegel (2017), in comparative backtesting, we are interested in the following null hypotheses:

$$H_0^-: \quad \text{Forecast } x_{1,t} \text{ predicts at least as well as } x_{2,t},$$
$$H_0^+: \quad \text{Forecast } x_{1,t} \text{ predicts at most as well as } x_{2,t}.$$

The forecast $x_{2,t}$ is used as a benchmark. If the hypothesis H_0^- is rejected, then $x_{2,t}$ is worse than $x_{1,t}$; if H_0^+ is rejected, $x_{1,t}$ is better than $x_{2,t}$. The error of the first kind for rejecting one of the two hypotheses, even though they are true, can be controlled by the level of significance. As in Nolde and Ziegel (2017), we define

$$\lambda = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\big[\mathrm{L}(x_{1,t}, Y_{t+1}) - \mathrm{L}(x_{2,t}, Y_{t+1})\big] = \mathbb{E}\big[\mathrm{L}(x_{1,t}, Y_{t+1}) - \mathrm{L}(x_{2,t}, Y_{t+1})\big]$$

(assuming first-order stationarity). Then dominance of $x_{1,t}$ over $x_{2,t}$ is equivalent to $\lambda \le 0$, and $x_{1,t}$ predicts at most as well as $x_{2,t}$ if $\lambda \ge 0$. Therefore, the comparative backtesting hypotheses can be reformulated as

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

Forecast equality can be tested with the so-called DM test (Diebold and Mariano (1995), Giacomini and White (2006), Diebold (2015)), which is based on normalized loss differences. Here, the test statistic is given by

$$S = \frac{\sqrt{n}\Delta_n \bar{L}}{\hat{\tau}},$$

where $\Delta_n \bar{L} = 1/n \sum_{t=1}^{n} L(x_{1,t}, y_{t+1}) - L(x_{2,t}, y_{t+1})$ and $\hat{\tau}^2$ is an estimator of the long-run asymptotic variance of the loss differences. One possible choice for $\hat{\tau}^2$ is

$$\hat{\tau}^2 = \begin{cases} \hat{\gamma}_0 & \text{if } h = 1, \\ \hat{\gamma}_0 + 2\sum_{j=1}^{h-1} \hat{\gamma}_j & \text{if } h \ge 2, \end{cases}$$

where $\hat{\gamma}_j$ denotes the lag j sample autocovariance of the sequence of loss differences (Gneiting and Ranjan (2011), Lerch et al. (2017)). Another possible choice (Diks, Panchenko and van Dijk (2011)) is $\hat{\tau}^2 = \hat{\gamma}_0 + 2\sum_{j=1}^{J}(1 - j/J)\hat{\gamma}_j$, where J is the largest integer less than or equal to $n^{1/4}$. As a compromise, we used $\hat{\tau}^2 = \hat{\gamma}_0 + 2\hat{\gamma}_1$. Under the null hypothesis of a vanishing expected loss difference and some further regularity conditions, the test statistic S is asymptotically standard normally distributed. Therefore, we obtain an asymptotic level-η test of H_0^+ if we reject the null hypothesis when $S \le \Phi^{-1}(\eta)$, and of H_0^- if we reject the null hypothesis when $S \ge \Phi^{-1}(1-\eta)$.

To evaluate the tests for a fixed significance level $\eta \in (0, 1)$, we use the following three-zone approach of Fissler, Ziegel and Gneiting (2016). If H_0^- is rejected at level η, we conclude that the forecast $x_{1,t}$ is worse than $x_{2,t}$, and we mark the result in red; similarly, if H_0^+ is rejected at level η, forecast $x_{1,t}$ is better than $x_{2,t}$, and we mark the result in green. Finally, if neither H_0^- nor H_0^+ can be rejected, the marking is yellow.

6. Simulations. In this section, we report some results of an extensive simulation study. Additional results of the simulations are contained in Section 1 of the Supplementary Material (Holzmann and Klar (2023)). First, in Section 6.1 we investigate proxies for the volatility, and then in Sections 6.2 and 6.3 turn to higher-order moments.

6.1. *Squared returns and realized variance.* In the first two sections, as data generating process (DGP) for the log returns we use a GARCH(1,1) process defined by

$$\sigma_t^2 = a_0 + a_1 r_{t-1}^2 + b\sigma_{t-1}^2, \quad r_t = \sigma_t \varepsilon_t,$$

where additionally

$$\varepsilon_t = \sum_{i=1}^{m} \varepsilon_{t,i}, \varepsilon_{t,i} = \mathcal{N}(0, 1/m), \quad i = 1, \dots, m, \tag{26}$$

and all $\varepsilon_{t,i}$ independent. Assuming that σ_t is constant on $(t-1, t]$, observed intraday returns are given by $r_{t,i} = \sigma_t \varepsilon_{t,i}$. We use $a_0 = 0.02, a_1 = 0.08, b = 0.85$, $m = 100$, and $m = 13$; the first is a typical range using 5-min returns, the latter corresponds to the use of half-hour returns at the New York Stock Exchange (NYSE). While this is certainly an oversimplified model for actual high-frequency data, it serves to illustrate the use of proxies, and is similar to the setting used in Patton (2011), Section 2.2. Real high-frequency data for log-returns of cryptocurrencies are analyzed in Section 7.

As the total length of the time series, we take $T = 1500$ and $T = 6000$. For these two time series, we generate rolling one-step-ahead forecasts of the conditional variance using a moving time window, with window length $T/3$, refitted every 10 time steps, for GARCH(1,1), ARCH(1), ARCH(2), and ARCH(7) models. Hence, for $T = 1500$, the fit is based on 500 values, and the DM tests use 1000 forecasts of each model. All computations are done in R (R Core Team (2021)) using the R packages `rugarch` (Ghalanos (2020)) and `fGarch` (Wuertz et al. (2020)).

To stabilize the results, the following figures show the means of the results of 500 replications for Figures 2 and 3, and of 50 replications for all figures after Figure 3 of the DM test. All figures use the three-zone approach described in the previous section with the following modification: we simultaneously show rejection of H_0^- at level 0.1, 0.05 and 0.01 by marking in light red, red, and dark red, respectively, marking in light green, green, and dark green signals rejection of H_0^+ at the three levels. Besides the forecasts from the different (G)ARCH models, we show the result for the optimal forecast, given by the true conditional volatilities. Each figure shows four plot matrices: in the left (right) column, the squared

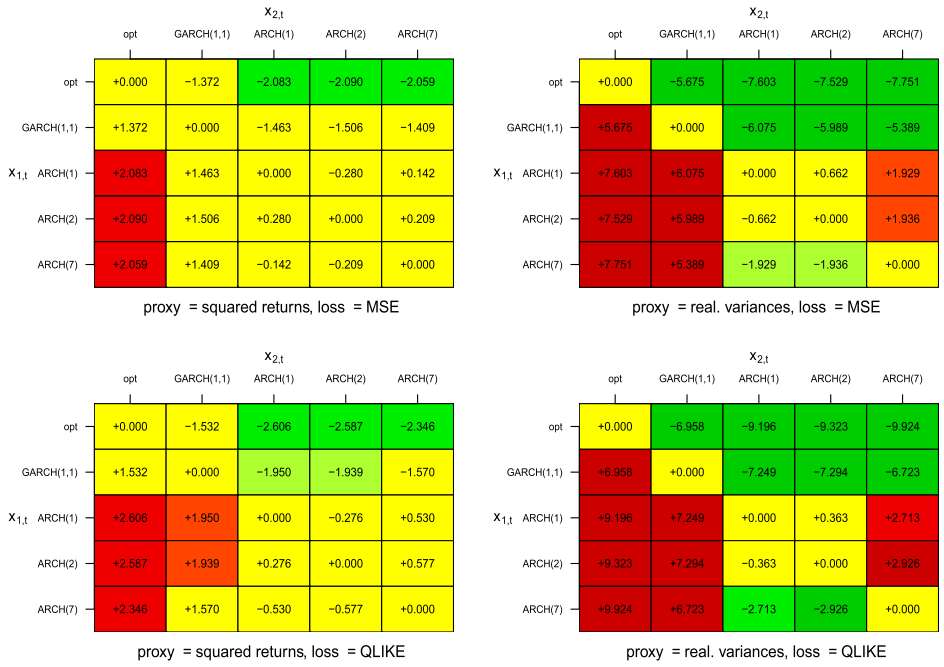

FIG. 2. *Results of DM tests* (*based on* 500 *replications*), *normal distribution*, *left*: *squared returns*, *right*: *realized variances*, $m = 100$, $T = 1500$.

returns (realized variances) are used as proxies. In the upper row, the loss function is the mean squared error $L(x, y) = (x - y)^2$, whereas in the lower row, the QLIKE loss function $\tilde{L}(x, y) = y/x - \log(y/x) - 1$ is used. These loss functions correspond to the choice of $\phi(y) = y^2$ and $\phi(y) = -\log(y)$ in equation (7). The power of the DM test and even the ranking of competing forecasts may depend on the particular choice of the strictly consistent loss function in case of misspecified forecasts; see Patton (2020) and Ehm et al. (2016) for a discussion. The QLIKE loss for the mean was proposed in Patton (2011) as a 0-homogeneous alternative to the more standard squared loss. It requires fewer moment assumptions, and a favorable performance for volatility forecasting was found in various empirical studies including Patton and Sheppard (2009).

Figure 2 shows the results of DM tests under normal innovations with $T = 1500$ and $m = 100$. The left panels show results for the squared returns r_t^2, the right panels for the realized variances $RV_t = \sum_{i=1}^{m} r_{t,i}^2$.

Let us first discuss the results shown in the lower-left panel, that is, for the QLIKE loss and using the squared returns as proxies. The second value in the left column, $+1.532$, is the average value of the DM test statistic comparing the forecast $x_{1,t}$ from the GARCH(1,1) model with the optimal forecast $x_{2,t}$, the true conditional volatilities. The positive value hints at the superiority of $x_{2,t}$, but the value is not statistically significant. The results are significant when comparing the ARCH(1), ARCH(2), and ARCH(7) model with the optimal forecast. Here, the red color indicates a significant rejection of H_0^- at the 0.05-level. The light red entries in the second column show that forecasts from the ARCH(1) and ARCH(2) models are worse than the forecasts from the GARCH(1,1) model (which is the true data generating process) at level 0.1, but not on the 0.05-level. The forecast from GARCH(7) is not significantly worse than the GARCH(1,1).

Now, let us turn to the lower-right panel with the realized variances as proxies. Here, all corresponding entries are marked in dark red, signalizing the rejection of H_0^- at level 0.01 in all cases. Hence, the power of the DM test is clearly higher by using realized variances

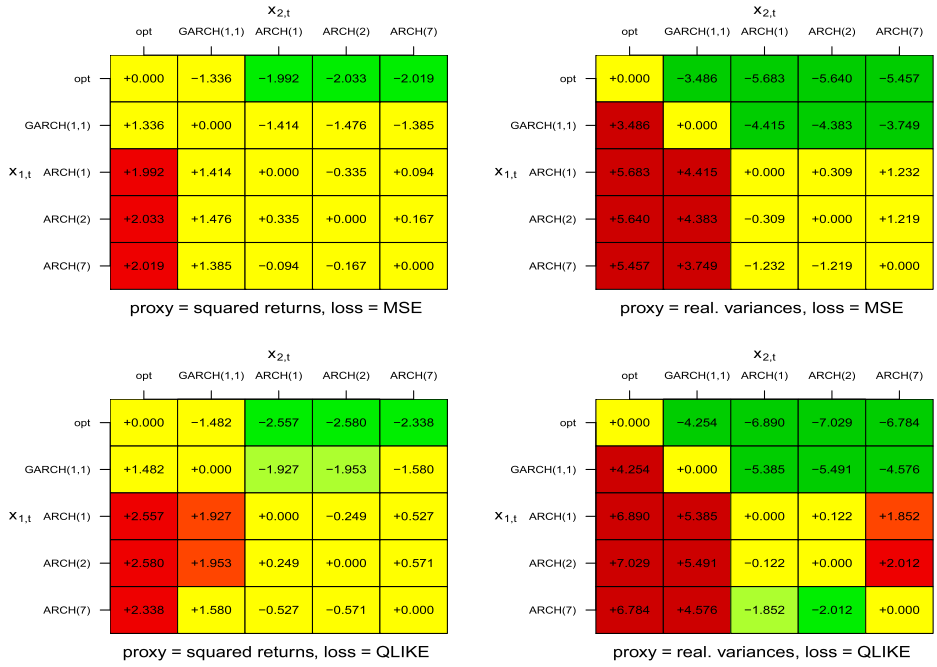

FIG. 3. *Results of DM tests* (*based on* 500 *replications*), *normal distribution*, *left*: *squared returns*, *right*: *realized variances*, $m = 13$, $T = 1500$.

compared to squared returns. Looking at the upper row, we see that the results for the MSE are similar from a qualitative point of view. However, there are fewer statistically significant entries compared to the QLIKE loss function. Hence, the latter allows for a sharper forecast evaluation in this example.

Figure 3 shows results from the same setting as Figure 2 apart from that we use $m = 13$, corresponding to the use of half-hourly returns, instead of $m = 100$. Hence, the results in the two left panels are the same as in Figure 2 (up to simulation error). The right panels are similar as in Figure 2, as well. A closer look shows that all entries have smaller absolute values, showing the decreasing power in differentiating forecasts.

We also replaced the normal distribution of the intraday innovations by centered skewed and long-tailed distributions. For this, we used the normal inverse Gaussian distribution $nig(\mu, \delta, \alpha, \beta)$ (Barndorff-Nielsen (1997)) with parameters

$$\alpha = 2, \qquad \beta = 1, \qquad \gamma = \sqrt{\alpha^2 - \beta^2}, \qquad \delta = \gamma^3/\alpha^2/m, \qquad \mu = -\delta\beta/\gamma.$$

This results in $\mathbb{E}[\varepsilon_{t,i}] = 0$ and $\mathrm{Var}(\varepsilon_{t,i}) = 1/m$. Since the class of nig distributions with fixed shape parameters α and β is closed under convolution, the distribution of $\varepsilon_t = \sum_{i=1}^{m} \varepsilon_{t,i}$ is given by $nig(m\mu, m\delta, \alpha, \beta)$, with $\mathbb{E}[\varepsilon_t] = 0$, $\mathrm{Var}(\varepsilon_t) = 1$, $\mathbb{E}[\varepsilon_t^3] = 1$, and $\mathbb{E}[\varepsilon_t^4] = 17/3$.

Figure 4 shows the results of the DM tests as in Figure 2, that is, for $m = 100$, using nig instead of normally distributed innovations. Again, the results are qualitatively comparable to the results in Figure 2, but the absolute values of the entries are generally smaller. Hence, the change in the distribution of the innovations has a negative effect on the power of the test. Note that this decrease of power is larger for the realized variances than for the squared returns. This can be explained by the fact that the skewness and kurtosis of the daily innovations are rather modest with values of 1 and 5.67, whereas the skewness and kurtosis of the intraday innovations are 10 and 269.8, respectively.

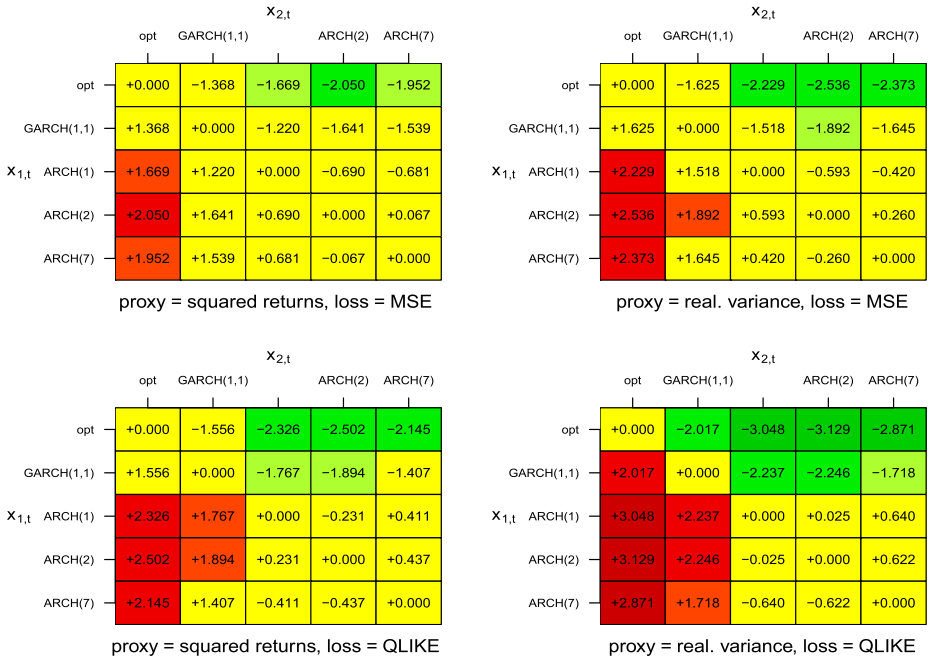

FIG. 4. *Results of DM tests* (*based on* 50 *replications*), *nig distribution*, *left*: *squared returns*, *right*: *realized variances*, $m = 100$, $T = 1500$.

6.2. *Higher moments*. The use of realized higher moments, skewness, and kurtosis to estimate and forecast returns has become quite standard in the literature. For example, Neuberger (2012) analyzed realized skewness and showed that high-frequency data can be used to provide more efficient estimates of the skewness in price changes over a period. Amaya et al. (2015) constructed measures of realized daily skewness and kurtosis based on intraday returns, and analyzed moment-based portfolios. Recently, Shen, Yao and Li (2018) discussed the explanatory power of higher realized moments.

6.2.1. *Third moment*. Assuming $\mathbb{E}[r_t|\mathcal{F}_{t-1}] = 0$, we are interested in the conditional third moment $\rho_t = \mathbb{E}[r_t^3|\mathcal{F}_{t-1}]$. Possible proxies for ρ_t are the cubed return r_t^3 and the realized third moment $RM(3)_t = \sum_{i=1}^m r_{t,i}^3$. We use the GARCH(1,1) model of Section 6.1, with the normal inverse Gaussian distribution for the innovations. Under this model, we obtain

$$\rho_t = \mathbb{E}\big[r_t^3|\mathcal{F}_{t-1}\big] = \mathbb{E}\big[\sigma_t^3\varepsilon_t^3|\mathcal{F}_{t-1}\big] = \sigma_t^3\mathbb{E}\big[\varepsilon_t^3\big],$$

$$\mathbb{E}\big[\varepsilon_t^3\big] = \mathbb{E}\left[\left(\sum_{i=1}^m \varepsilon_{t,i}\right)^3\right]$$

$$= \mathbb{E}\left[\sum_i \varepsilon_{t,i}^3 + 3\sum_{i<j}\varepsilon_{t,i}^2\varepsilon_{t,j} + 6\sum_{i<j<k}\varepsilon_{t,i}\varepsilon_{t,j}\varepsilon_{t,k}\right] = \sum_{i=1}^m \mathbb{E}\big[\varepsilon_{t,i}^3\big].$$

Since

$$\mathbb{E}\big[RM(3)_t|\mathcal{F}_{t-1}\big] = \sigma_t^3\sum_{i=1}^m \mathbb{E}\big[\varepsilon_{t,i}^3\big] = \sigma_t^3\mathbb{E}\big[\varepsilon_t^3\big],$$

$RM(3)_t$ is an unbiased estimator of ρ_t. As a forecast of ρ_t, we use $\tilde{\sigma}_t^3\mathbb{E}[\varepsilon_t^3]$, where $\tilde{\sigma}_t$ denotes the one-step ahead forecast of σ_t from the different (G)ARCH models.

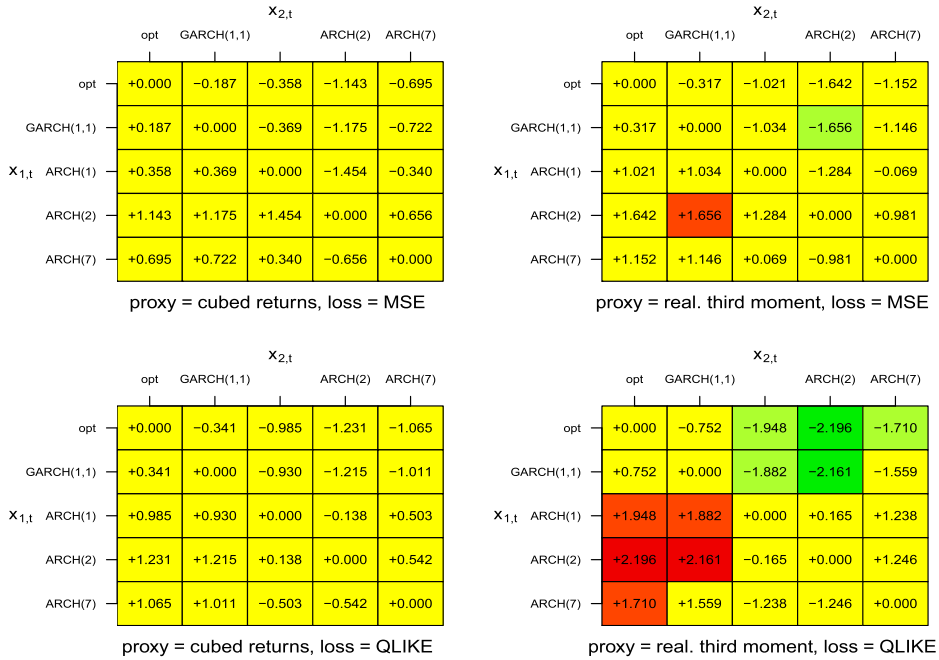

FIG. 5. *Results of DM tests* (*based on* 50 *replications*), *nig distribution*, *left*: *cubed returns*, *right*: *realized third moment*, $m = 13$, $T = 6000$.

Figure 5 shows the results of DM tests under $nig(\mu, \delta, 2, 1)$ innovations, with μ, δ chosen such that $\mathbb{E}[\varepsilon_{t,i}] = 0$, $\mathrm{Var}(\varepsilon_{t,i}) = m^{-1/2}$. Skewness and kurtosis of the intraday innovations are 3.61 and 37.67, respectively, compared to the values 1 and 5.67 of the daily innovations. Here, total length of the simulated time series is $T = 6000$, and we use $m = 13$, that is, half-hourly returns. The left panels show the results for the cubed returns r_t^3, the right panels for the realized third moment $RM(3)_t = \sum_{i=1}^m r_{t,i}^3$.

At first glance, the results seem to be rather different from the corresponding ones for the volatility, since the number of significant entries is much lower (cf. Figure 3). But they go in the same direction: use of the realized moments increases the power of the DM test when the optimal forecast competes against the other models, or when the true data generating process is compared with ARCH models.

We have also used $T = 1500$ in the simulations; the results (not shown) go in the same direction, but none of the values are statistically significant, even at the 0.1-level.

6.2.2. *Fourth moment*. Here, we are interested in the conditional fourth moment $\tau_t = \mathbb{E}[r_t^4|\mathcal{F}_{t-1}]$. Again, we use the GARCH(1,1) model as in Section 6.1, and obtain

$$\mathbb{E}[\varepsilon_t^4] = \sum_i \mathbb{E}[\varepsilon_{t,i}^4] + 6\sum_{i<j} \mathbb{E}[\varepsilon_{t,i}^2 \varepsilon_{t,j}^2] = \sum_i \mathbb{E}[\varepsilon_{t,i}^4] + 3m(m-1)\big(\mathbb{E}[\varepsilon_{t,1}^2]\big)^2.$$

Hence, unbiased proxies for τ_t are r_t^4 and the realized corrected fourth moment

$$cRM(4)_t = \sum_{i=1}^m r_{t,i}^4 + 6\sum_{i<j} r_{t,i}^2 r_{t,j}^2.$$

As a forecast of τ_t, we use $\tilde{\sigma}_t^4 \mathbb{E}[\varepsilon_t^4]$.

The left and right panels of Figure 6 show the results of the DM tests, using r_t^4 and the realized corrected fourth moment as proxies, respectively. The innovations are normally distributed; further, $T = 1500$ and $m = 13$. The general picture resembles strongly the results

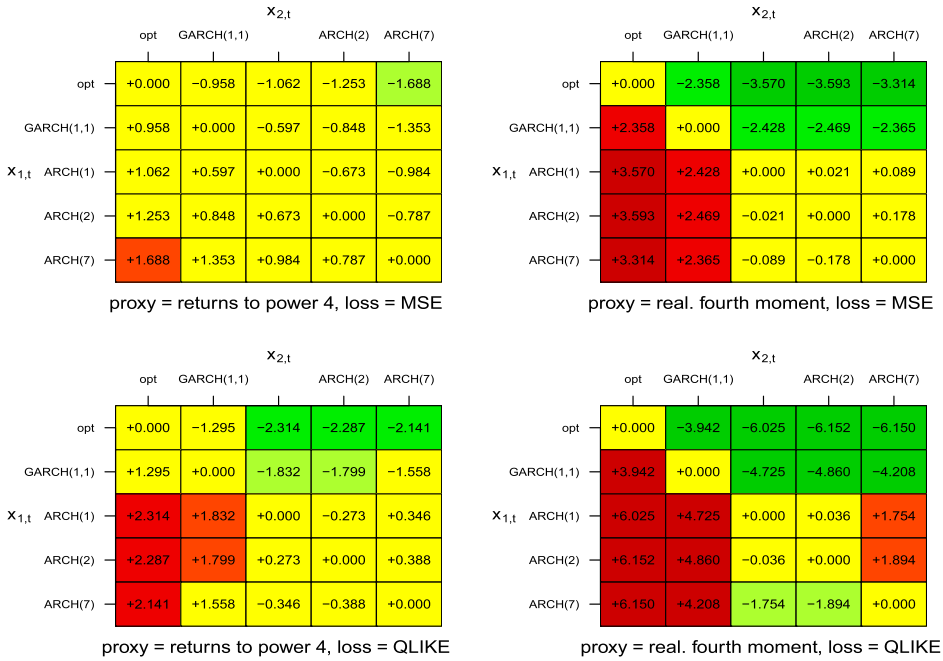

FIG. 6. *Results of DM tests* (*based on* 50 *replications*), *normal distribution*, *left*: *returns to the power* 4, *right*: *realized corrected fourth moment*, $m = 13$, $T = 1500$.

of the volatility forecasts in Figure 3, and all conclusions also apply here, even though the actual entries are a bit smaller.

When replacing the normal by the nig innovations, the power of the DM test decreases strongly (cf. Figure A.2 in the Supplementary Material (Holzmann and Klar (2023))). On the other hand, the entries are somewhat larger as in forecasting the third moment (with $T = 1500$). Here, at least a few values are significant on the 0.1-level.

6.3. *An apARCH model for the fourth moment.* Instead of modeling the volatility, and computing higher moments under this process, it is also possible to use suitable models for higher moments directly. Harvey and Siddique (1999, 2000), for example, considered an autoregressive model for conditional skewness. Lambert and Laurent (2002) used the asymmetric power (G)ARCH or APARCH model of Ding, Granger and Engle (1993) to describe dynamics in skewed location-scale distributions. Brooks et al. (2005) used both separate and joint GARCH models for conditional variance and conditional kurtosis, whereas Lau (2015) modeled (standardized) realized moments by an exponentially weighted moving average.

Hence, in this section, we model the fourth moment directly by an asymmetric power ARCH (apARCH) process (Ding, Granger and Engle (1993)). Specifically, the log returns follow an apARCH(1,1) model with $\delta = 4$,

$$\sigma_t^4 = \omega + \alpha r_{t-1}^4 + \beta \sigma_{t-1}^4, \quad r_t = \sigma_t \varepsilon_t,$$

where $\varepsilon_t = \sum_{i=1}^m \varepsilon_{t,i}$, $\varepsilon_{t,i} = \mathcal{N}(0, 1/m)$ for $i = 1, \dots, m$, and all $\varepsilon_{t,i}$ are independent. Assuming again that σ_t is constant on $(t-1, t]$, intraday returns are given by $r_{t,i} = \sigma_t \varepsilon_{t,i}$. We use $\omega = 0.02, \alpha = 0.08, \beta = 0.75$ such that the unconditional variance is

$$\sigma^2 = \left(\frac{\omega}{1 - E(\varepsilon_1^4)\alpha - \beta} \right)^{2/\delta} = \sqrt{2}.$$

Further, $m = 100$ and $T = 1500$. As in the last section, unbiased proxies for $\tau_t = \mathbb{E}[r_t^4 | \mathcal{F}_{t-1}]$ are r_t^4 and $cRM(4)_t$. As a forecast of τ_t, we use $\tilde{\sigma}_t^4 \mathbb{E}[\varepsilon_t^4]$, where $\tilde{\sigma}_t^4$ denotes the one-step

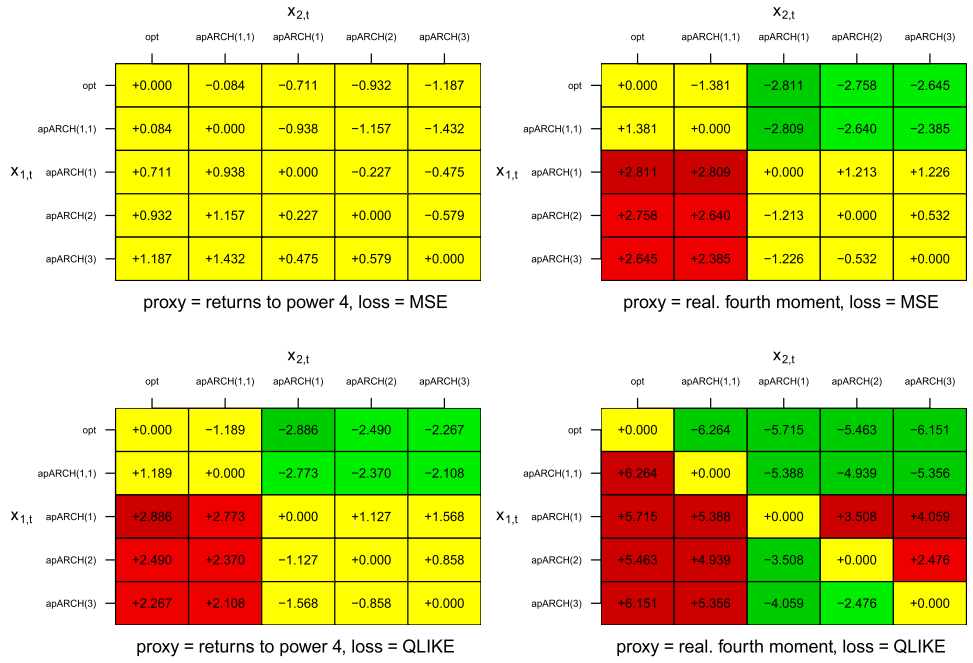

FIG. 7. *Results of DM tests* (*based on* 50 *replications*), *apARCH process with exponent* 4, *normal distribution*, *left*: *returns to the power* 4, *right*: *realized corrected fourth moment*, $m = 100$, $T = 1500$.

ahead forecast of σ_t^4 from the different apARCH models, namely apARCH(1,1), apARCH(1), apARCH(2), and apARCH(3).

The left and right panels of Figure 7 show the results of the DM tests for the apARCH process with exponent 4, with r_t^4 and the realized corrected fourth moment, respectively, as proxies.

The visual comparison of the upper-left and lower-right panels is striking; in the latter, each result is significant, whereas the former shows no significant entries. Hence, using high-frequency data and a suitable loss function results in a highly improved forecast evaluation. Generally, the results are quite similar to the results for the fourth moment based on the GARCH process in Figure 6.

To sum up the results of the simulations, it has become obvious that using high-frequency data for the proxies improves the forecast evaluation in each example. In most cases, the effect is substantial. There is also an effect of the choice of the loss function: the power of the DM test improves when using the QLIKE loss compared to the MSE loss function.

7. Log returns of cryptocurrencies. Many GARCH and GARCH-type models have been used for modeling and predicting the volatility of cryptocurrencies (Katsiampa (2017), Naimy and Hayek (2018), Chu et al. (2017)), and there is no general agreement which model is the best choice. Katsiampa (2017), Naimy and Hayek (2018), and Gyamerah (2019) advocate the use of the component GARCH (CGARCH), the exponential GARCH (EGARCH) and the threshhold GARCH (TGARCH) model, respectively. Chu et al. (2017) conclude that the standard GARCH (SGARCH), the integrated GARCH (IGARCH), and the Glosten–Jagannathan–Runkle GARCH (GJRGARCH) are preferable depending on the cryptocurrency. An overview over this and related literature can be found in Naimy et al. (2021), who favor CGARCH, GJR-GARCH, APARCH, and TGARCH. Catania, Grassi and Ravazzolo (2018) advocate the use of a score driven model with conditional generalized hyperbolic skew student's-t innovations for predicting the conditional volatility.

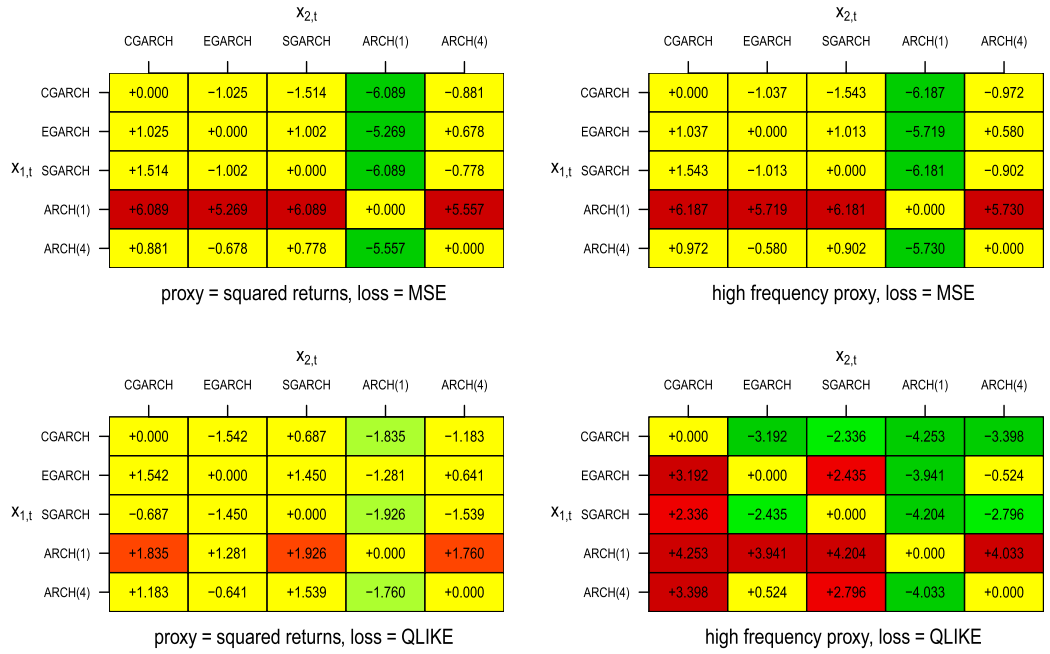

FIG. 8. *Results of DM tests for ETH log returns from May* 16, 2018, *to October* 27, 2021, *left*: *squared returns*, *right*: *high frequency proxy.*

In this section, we consider log returns of three cryptocurrencies, namely Bitcoin (BTC), already used in Section 2.1, Ethereum (ETH), and Ripple (XRP). According to coinmarketcap.com, BTC and ETH are the cryptocurrencies with the highest and second highest market capitalization; XRP is number seven in the list. BTC is one of the oldest cryptocurrencies, existing since 2008, whereas ETH and XRP were released in 2013 and 2012. All three cryptocurrencies are traded in many cryptocurrency exchanges like Binance, FTX, or Bitstamp. We use hourly observations from May 16, 2018, to October 27, 2021, with sample size 30264, which corresponds to 1261 days. All prices are closing values in U.S. dollars of the Bitstamp Exchange obtained from cryptodatadownload.com. Returns are estimated by taking logarithmic differences. Figures A.4 and A.5 in the Supplementary Material (Holzmann and Klar (2023)) show plots of the log returns and the autocorrelation functions of log returns for the three cryptocurrencies.

Similarly, as in Section 6, we consider a standard GARCH(1,1) model and ARCH models of order 1 and 4. As more sophisticated models, we also employ the EGARCH model of Nelson (1991) and the CGARCH model of Lee and Engle (1999), both with $p = q = 1$ and normal innovations. We model the conditional mean by a constant value in all cases. Since the variance is nonelicitable, we aim at predicting the conditional second moment, using either squared returns or the high frequency proxy $\sum_{i=1}^{24} r_{t,i}^2$. One-step-ahead forecasts use a moving time window with length $\lfloor T/3 \rfloor = 420$, refitted every time step. Figure 1 in Section 2.1 and Figures 8 and 9 show the results of DM tests for the log returns of Bitcoin, Ethereum, and Ripple, respectively.

The general observation is that the GARCH-type models outperform the two ARCH models. The GARCH-type processes are comparable, with the CGARCH dominating the EGARCH model for BTC in case of the QLIKE loss and high frequency proxy. For QLIKE loss and high frequency proxy, CGARCH outperforms all other models, followed by SGARCH.

With QLIKE loss, the power of the DM tests is considerably higher by using the high frequency proxy compared to squared returns. Moreover, for BTC and ETH, there is an in-

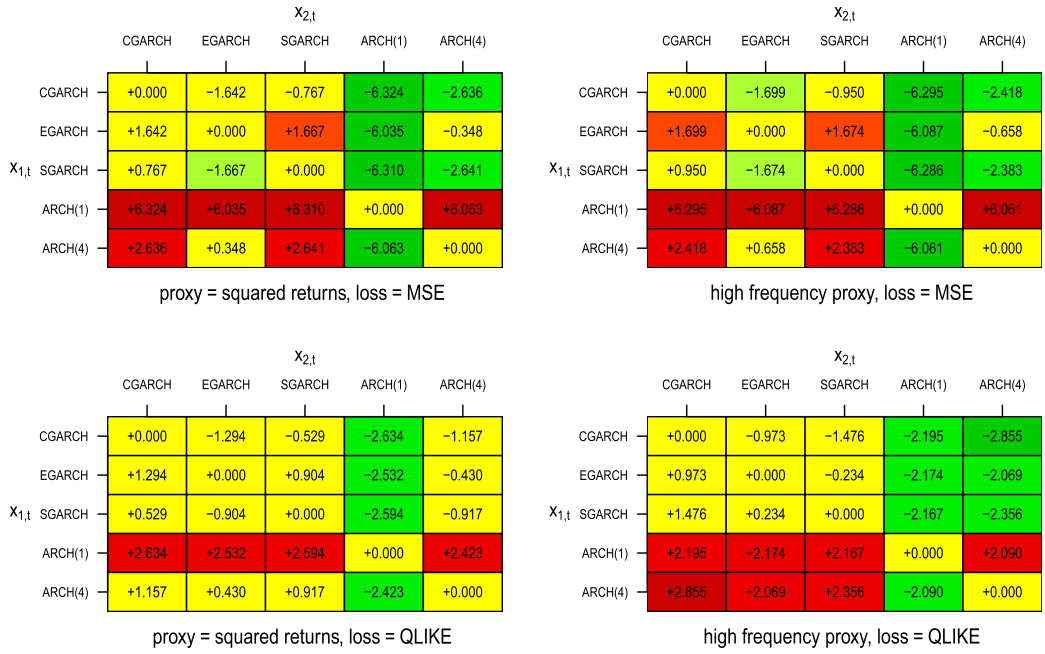

FIG. 9. *Results of DM tests for XRP log returns from May* 16, 2018, *to October* 27, 2021, *left*: *squared returns*, *right*: *high frequency proxy.*

crease in power using the QLIKE loss function compared to MSE; somewhat surprisingly, the reverse holds for XRP. Hence, apart from the latter observation, the results corroborate the findings of the simulations.

8. Concluding remarks. Our perspective, say in the Theorem 2, is that Y_{t+1} is observed at time $t+1$, but that in addition a proxy $\tilde{Y}_{t+1}$ is also observed or can be computed, which is more informative about the functional. Hence, $\tilde{Y}_{t+1}$ can be used to increase the power of tests for forecast dominance. The theoretical results are supported by both the simulations and the real data example: The use of high-frequency data for the proxies generally improves the forecast evaluation. This effect overlaps with the impact of the choice of the loss function, which can also be quite high. Empirically, the power of the DM test improves when using the QLIKE loss compared to the MSE loss function, although no theoretical results seem to be available in this direction.

Another perspective which is pursued, for example, in Hoga and Dimitriadis (2022) for mean forecasts of US GDP growth, by Li and Patton (2018) for forecasting integrated volatility in high-frequency finance, and by Kleen (2021) for probabilistic forecasts is that the variable of interest Y_{t+1} is actually not observed but latent, and only proxies of Y_{t+1} with additional measurement error can be observed. The problem is to quantify the effect of measurement error on forecast evaluation. An interesting issue in the context of probabilistic forecasting would thus be to investigate whether proxies can also be used, as in our setting, to improve forecast evaluation.

When moving beyond moments and ratios of moments and when considering general functionals T, one always has the following: If L is strictly consistent for T, and $\tilde{Y}_{t+1}$ is a proxy for T of Y_{t+1} in the sense that the conditional functionals

$$T\big(F_{Y_{t+1}|\mathcal{F}_t}(\omega,\cdot)\big) = T\big(F_{\tilde{Y}_{t+1}|\mathcal{F}_t}(\omega,\cdot)\big) \tag{27}$$

coincide, then for $\mathcal{F}_t$-measurable Z_t,

$$\mathbb{E}\big[\mathrm{L}\big(T\big(F_{Y_{t+1}|\mathcal{F}_t}(\omega,\cdot)\big),\tilde{Y}_{t+1}\big)|\mathcal{F}_t\big](\omega) \le \mathbb{E}\big[\mathrm{L}(Z_t,\tilde{Y}_{t+1})|\mathcal{F}_t\big](\omega) \quad \text{for } \mathbb{P}\text{-a.e. } \omega\in\Omega, \tag{28}$$

with equality almost surely if and only if $T(F_{Y_{t+1}|\mathcal{F}_t}(\omega,\cdot)) = Z_t$ $\mathbb{P}$-almost surely. Indeed, this is simply (3) stated for $\tilde{Y}_{t+1}$ by observing (27). (28) implies that if we do not take into account comparing *two* possibly misspecified forecasts, then we can always replace the variable of interest Y_{t+1} by a proxy $\tilde{Y}_{t+1}$, which satisfies (27).

However, Hoga and Dimitriadis (2022) show in their Proposition 3 that for quantile scores, this cannot be extended to comparing misspecified forecasts: If the difference of conditional loss differences vanishes, then the conditional distributions of Y_{t+1} and $\tilde{Y}_{t+1}$ must coincide. It would be of interest to investigate if this negative result is more pervasive and applies to further functionals such as expectiles.

Acknowledgments. We would like to thank Andrew Patton and Tilmann Gneiting for pointers to the literature, and in particular for bringing the paper by Hoga and Dimitriadis (2022) to our attention. Then we would like to thank Timo Dimitriadis for providing various general and detailed, helpful comments on the paper. The authors are grateful for the suggestions and comments of two anonymous reviewers, which helped to improve the paper.

SUPPLEMENTARY MATERIAL

Supplementary material (DOI: 10.1214/22-AOAS1716SUPP; .pdf). Additional simulation results and material for the real data analysis are available as supplementary material.

REFERENCES

AMAYA, D., CHRISTOFFERSEN, P., JACOBS, K. and VASQUEZ, A. (2015). Does realized skewness predict the cross-section of equity returns? *J. Financ. Econ.* **118** 135–167.

BARNDORFF-NIELSEN, O. E. (1997). Normal inverse Gaussian distributions and stochastic volatility modelling. *Scand. J. Stat.* **24** 1–13. MR1436619 https://doi.org/10.1111/1467-9469.00045

BROOKS, C., BURKE, S. P., HERAVI, S. and PERSAND, G. (2005). Autoregressive conditional kurtosis. *J. Financ. Econom.* **3** 399–421.

CATANIA, L., GRASSI, S. and RAVAZZOLO, F. (2018). Predicting the volatility of cryptocurrency time-series. In *Mathematical and Statistical Methods for Actuarial Sciences and Finance* (M. Corazza, M. Durbán, A. Grané, C. Perna and M. Sibillo, eds.). Springer, Berlin.

CHU, J., CHAN, S., NADARAJAH, S. and OSTERRIEDER, J. (2017). Garch modelling of cryptocurrencies. *J. Financ. Risk Manag.* **10**.

CORSI, F. (2009). A simple approximate long-memory model of realized volatility. *J. Financ. Econom.* **7** 174–196.

DIEBOLD, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *J. Bus. Econom. Statist.* **33** 1–9. MR3303732 https://doi.org/10.1080/07350015.2014.983236

DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econom. Statist.* **33** 253–263.

DIKS, C., PANCHENKO, V. and VAN DIJK, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *J. Econometrics* **163** 215–230. MR2812867 https://doi.org/10.1016/j.jeconom.2011.04.001

DING, Z., GRANGER, C. and ENGLE, R. (1993). A long memory property of stock market returns and a new model. *J. Empir. Finance* **83** 83–106.

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 https://doi.org/10.1111/rssb.12154

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

GHALANOS, A. (2020). rugarch: Univariate GARCH models. R package version 1.4-4.

GIACOMINI, R. and WHITE, H. (2006). Tests of conditional predictive ability. *Econometrica* **74** 1545–1578. MR2268409 https://doi.org/10.1111/j.1468-0262.2006.00718.x

GNEITING, T. (2011). Making and evaluating point forecasts. *J. Amer. Statist. Assoc.* **106** 746–762. MR2847988 https://doi.org/10.1198/jasa.2011.r10138

GNEITING, T. and RANJAN, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econom. Statist.* **29** 411–422. MR2848512 https://doi.org/10.1198/jbes.2010.08110

GYAMERAH, S. (2019). Modelling the volatility of bitcoin returns using GARCH models. *Quantitative Finance and Economics* **3** 739–53.

HANSEN, P. R. and LUNDE, A. (2006). Consistent ranking of volatility models. *J. Econometrics* **131** 97–121. MR2275996 https://doi.org/10.1016/j.jeconom.2005.01.005

HARVEY, C. R. and SIDDIQUE, A. (1999). Autoregressive conditional skewness. *J. Financ. Quant. Anal.* **34** 465–487.

HARVEY, C. R. and SIDDIQUE, A. (2000). Conditional skewness in asset pricing tests. *J. Finance* **55** 1263–1295.

HOGA, Y. and DIMITRIADIS, T. (2022). On testing equal conditional predictive ability under measurement error. *J. Bus. Econom. Statist.* **0** 1–13.

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 https://doi.org/10.1214/13-AOAS709

HOLZMANN, H. and KLAR, B. (2023). Supplement to "Using proxies to improve forecast evaluation." https://doi.org/10.1214/22-AOAS1716SUPP

KATSIAMPA, P. (2017). Volatility estimation for Bitcoin: A comparison of GARCH models. *Econom. Lett.* **158** 3–6. MR3681256 https://doi.org/10.1016/j.econlet.2017.06.023

KLEEN, O. (2021). Measurement error sensitivity of loss functions for distribution forecasts. Available at SSRN 3476461.

KOOPMAN, S. J., JUNGBACKER, B. and HOL, E. (2005). Forecasting daily variability of the s&p 100 stock index using historical, realised and implied volatility measurements. *J. Empir. Finance* **12** 445–475.

LAMBERT, P. and LAURENT, S. (2002). Modeling skewness dynamics in series of financial data using skewed location-scale distributions. Working Paper, Université Catholique de Louvain and Université de Liège.

LAU, C. (2015). A simple normal inverse Gaussian-type approach to calculate value-at-risk based on realized moments. *J. Risk* **17** 1–18.

LAURENT, S., ROMBOUTS, J. V. K. and VIOLANTE, F. (2013). On loss functions and ranking forecasting performances of multivariate volatility models. *J. Econometrics* **173** 1–10. MR3019678 https://doi.org/10.1016/j.jeconom.2012.08.004

LEE, G. J. and ENGLE, R. F. (1999). A permanent and transitory component model of stock return volatility. In *Cointegration, Causality and Forecasting: A Festschrift in Honor of Clive W. J. Granger* 980–996.

LERCH, S., THORARINSDOTTIR, T. L., RAVAZZOLO, F. and GNEITING, T. (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statist. Sci.* **32** 106–127. MR3634309 https://doi.org/10.1214/16-STS588

LI, J. and PATTON, A. J. (2018). Asymptotic inference about predictive accuracy using high frequency data. *J. Econometrics* **203** 223–240. MR3770823 https://doi.org/10.1016/j.jeconom.2017.10.005

NAIMY, V. and HAYEK, M. (2018). Modelling and predicting the bitcoin volatility using garch models. *Int. J. Math. Model. Numer. Optim.* **8** 197–215.

NAIMY, V., HADDAD, O., FERNÁNDEZ-AVILÉS, G. and EL KHOURY, R. (2021). The predictive capacity of GARCH-type models in measuring the volatility of crypto and world currencies. *PLoS ONE* **16** e0245904.

NELSON, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* **59** 347–370. MR1097532 https://doi.org/10.2307/2938260

NEUBERGER, A. (2012). Realized skewness. *Rev. Financ. Stud.* **25** 3423–3455.

NOLDE, N. and ZIEGEL, J. F. (2017). Elicitability and backtesting: Perspectives for banking regulation. *Ann. Appl. Stat.* **11** 1833–1874. MR3743276 https://doi.org/10.1214/17-AOAS1041

PATTON, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *J. Econometrics* **160** 246–256. MR2745881 https://doi.org/10.1016/j.jeconom.2010.03.034

PATTON, A. J. (2020). Comparing possibly misspecified forecasts. *J. Bus. Econom. Statist.* **38** 796–809. MR4154889 https://doi.org/10.1080/07350015.2019.1585256

PATTON, A. J. and SHEPPARD, K. (2009). Evaluating volatility and correlation forecasts. In *Handbook of Financial Time Series* 801–838. Springer, Berlin.

R CORE TEAM (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

SAVAGE, L. J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66** 783–801. MR0331571

SHEN, K., YAO, J. and LI, W. K. (2018). On the surprising explanatory power of higher realized moments in practice. *Stat. Interface* **11** 153–168. MR3690806 https://doi.org/10.4310/SII.2018.v11.n1.a13

STEINWART, I., PASIN, C., WILLIAMSON, R. and ZHANG, S. (2014). Elicitation and identification of properties. In *Conference on Learning Theory* 482–526. PMLR.

WUERTZ, D., SETZ, T., CHALABI, Y., BOUDT, C., CHAUSSE, P. and MIKLOVAC, M. (2020). fGarch: Rmetrics—Autoregressive Conditional Heteroskedastic Modelling. R package version 3042.83.2.