Model risk of volatility models

Model Risk of Volatility Models*

Emese Lazar†, Ning Zhang‡

Abstract A new model risk measure and estimation methodology based on loss functions is proposed in order to evaluate the accuracy of volatility models. The reliability of the proposed estimation has been verified via simulations and the estimates provide a reasonable fit to the true model risk measure. An empirical analysis based on several assets is undertaken to identify the models most affected by model risk, and concludes that the accuracy of volatility models can be improved by adjusting variance forecasts for model risk. The results indicate that after crisis situations, model risk increases especially for badly fitting volatility models.

Keywords: Model Risk, Scoring Functions, Volatility Forecast

JEL Classification: G17, G32, C22, C52, C58.

1. Introduction

Volatility forecasts often constitute a significant input in many financial applications, for example in risk estimation and investment decision-making. If the volatility forecast is inaccurate, then the implications can be widespread (Green and Figlewski, 1999). The existing extensive volatility modeling literature includes the family of autoregressive conditional heteroscedasticity (ARCH) models, stochastic volatility models as well as volatility models based on realized data, in a univariate or multivariate setting (see a detailed overview of volatility models in Bauwens et al., 2012). From the regulators' perspective (Federal Reserve, 2011), the term model refers to a quantitative approach or system that processes inputs and produces quantitative estimates. Consequently, the use of financial econometric techniques for volatility forecasting invariably presents model risk. This is defined as 'the potential for adverse consequences from decisions based on

*Computer codes are available from `https://gitfront.io/r/user-6451393/zdG7mbsRd1M4/model-risk-vol-models/`.

†Emese Lazar, ICMA Centre, Henley Business School, University of Reading, RG6 6BA, UK; e.lazar@icmacentre.ac.uk

‡Correspondence to: Ning Zhang, SILC Business School, Shanghai University, 201800, China; ning_zhang@shu.edu.cn

incorrect or misused model outputs and reports' (Federal Reserve, 2011), but there is no consensus on the exact definition of model risk in the current literature. Whilst some interpret it as the model error (relating to the difference between model output and true or realized value), others interpret it as the uncertainty about model error (Aven, 2016). In this paper, we define model risk as model error, because this is closely linked to the expected level of losses incurred if the model is used for volatility forecasting. For different applications, there will be a (not necessarily linear) relationship between losses and volatility model error, and often modelling this relationship introduces an extra layer of model risk. For example, the CAPM model assumes a linear relationship between returns and volatility; also, many risk models assume a linear relationship between Value-at-Risk (VaR, as a loss) and volatility, so a model error in volatility forecasting directly translates into a loss. However, many applications would have a non-linear relationship between volatility model error and losses (see Green and Figlewski, 1999). To keep our framework general and to measure pure volatility model risk, unaffected by the use of other models (e.g. pricing models), we quantify model risk as model error. In the aftermath of the 2008 global financial crisis, the Federal Reserve (2011) raises awareness of assessing model risk and so does the European Banking Authority (2014). We contribute to the literature by proposing a generic model risk measurement framework and estimating the model risk associated with two main sources of model risk: parameter estimation risk and model misspecification risk, both considered in Glasserman and Xu (2014) and Gourieroux and Monfort (2021).

Any volatility model exposes itself to the model risk associated with the distance from the model variance estimates to the true variance. However, the true variance (denoted by σ^2) is unobservable in practice, as discussed in Hansen and Lunde (2006) and Patton (2011), so the evaluation of the accuracy of volatility models is non-trivial. This can be addressed by using a conditionally unbiased variance estimator of the true conditional variance (hereafter, also called the variance proxy and denoted by $\hat{\sigma}^2$), namely the daily squared return, the realized variance, or the range-based variance to name the main ones (see Alizadeh et al., 2002; Barndorff-Nielsen and Shephard, 2002, Andersen et al., 2003 and Dette et al., 2022).

One strand of the volatility forecasting literature focuses on the accuracy of models taken in isolation. A simple and well-known approach to evaluate the accuracy of a single volatility model is the Mincer and Zarnowitz (1969) (MZ) regression (see more details in Section 5). The R^2 of the regression equation is considered as a criterion for the accuracy (efficiency) of the volatility forecasting model. A second strand, focusing on model evaluation, considers model comparisons based on scoring (loss) functions. The pairwise comparisons between two competing forecasts (see Diebold and Mariano, 1995, West, 1996, as well as a general discussion in Giacomini and White, 2006) and the multiple comparisons for a set of volatility models (e.g., Hansen and Lunde, 2005, and Hansen et al., 2011) have been well-documented. The main drawback of these tests is that the noisy variance proxy may distort the results, as argued by Hansen and Lunde (2006) and Patton (2011). To solve this, Patton (2011) proposes a class of robust and homogeneous scoring functions for the volatility, which leads to a ranking of competing models which is invariant to the choice of variance proxy. Within the proposed family, the mean squared error (MSE) and quasi-likelihood (QLIKE) scoring functions are widely accepted for the evaluation of volatility models as in Forsberg and Ghysels (2007), Bauwens et al. (2012), Engle and Siriwardane (2018) and others.

Although extensive studies of volatility forecast comparisons have been conducted based on the average loss, or distance between the estimated variances of competing models and variance proxies (e.g., Patton, 2011 and Hansen and Lunde, 2005), much less is known about the exact magnitude of model risk of the volatility forecast of individual models. Alternatively, concerning the uncertainty of point forecasts of volatilities, Pascual et al. (2006) apply a bootstrap procedure to compute the volatility prediction intervals as a measure of model risk which only allows for parameter estimation risk in a univariate setting; further, Fresoli and Ruiz (2016) measure the model risk of multivariate volatility models in a similar vein; Bollerslev et al. (2016) and Takahashi et al. (2021) improve on the forecasting models based on realized volatility. However, they do not specifically estimate model risk, nor consider the effects of parameter estimation risk and model misspecification risk as such. Recently, considerable attention has been devoted to the assessment of model risk of quantile-based risk measures like Value-at-Risk and Conditional Value-at-Risk (which are also unobservable in practice); this can be done in two ways: the first approach measures model risk related to the difference between the estimated risk and a reference risk estimate such as the upper or lower bound of risk estimates over a set of risk models as in Barrieu and Scandolo (2015) and Blasques et al. (2021), or as the ratio of the maximum risk forecast over the minimum within a set of models (Danielsson et al., 2016); the second approach measures model risk of a single risk model as the difference between the estimated risk measure and the improved risk measure that passes given backtests, see e.g. Boucher et al. (2014). Instead of computing model risk based on a subjective reference model, we measure the model risk of volatility models taken individually.

Our model risk estimation framework contributes to the literature in several ways. First, we expand the model risk literature by proposing a model risk measure for volatility models; we investigate the effect of using different objective functions (which are related to the choice of loss functions, i.e. MSE and QLIKE loss functions) and different variance proxies (i.e. the squared return and the realized variance) in computing model risk; and we develop an estimation methodology for model risk, considering different lengths of *optimization windows* and *model risk evaluation windows*. Specifically, for a given univariate volatility model we estimate model risk as the average distance between the raw and improved variance forecasts over a model risk evaluation window, in which the improved variance forecasts are obtained by minimizing the expected score of a given robust scoring function (MSE or QLIKE) over an optimization window.

Second, via Monte Carlo simulations we show the effectiveness of our proposed model risk estimation framework by comparing different model risk estimates with the true model risk across different lengths of optimization windows and model risk evaluation windows. Our results show that the proposed QLIKE-based model risk estimate, based on *additive adjustments* to the variance forecasts, is a good approximation of true model risk according to several measures of similarity. We find that the proposed model risk estimate has a correlation of at least 0.88 with the true model risk measure across the model set considered.

Third, considering the desirable coherence properties (Artzner et al., 1999) of a measure of risk for our proposed QLIKE-based model risk measure, we find that all desirable properties, specifically the monotonicity, positive homogeneity and translation invariance properties, are satisfied (except for subadditivity which is not required). As such, the proposed measure of model risk can be considered from a regulatory perspective.

Fourth, in an empirical study we compute the proposed QLIKE-based model risk estimates using different variance proxies (the squared return and the realized variance) for different asset classes, showing that all else being equal, the level of the the QLIKE-based model risk using the realized variance proxy is generally higher than that using the squared return proxy in small samples. The model risk of volatility models increases when the market volatility increases, which is in line with the result of Danielsson et al. (2016). We find that model risk has a negative effect on the predictive accuracy of volatility models, because the values of R^2 of the MZ regressions increase after adjusting the variance forecasts for model risk. We also disentangle the model risk of volatility models into *parameter estimation risk* and *model misspecification risk*, and conclude that model misspecification risk generally plays a more dominant role than parameter estimation risk.

The rest of the paper proceeds as follows: Section 2 introduces our proposed measure of model risk and estimation methodology based on the MSE and QLIKE loss functions; Section 3 justifies the introduced model risk estimates via simulations and compares different subjective choices; Section 4 examines the desirable coherence properties of our proposed measure; Section 5 applies the QLIKE-based model risk measure to different asset classes and distinguishes between different sources of model risk; Section 6 considers an alternative measure of model risk and Section 7 concludes.

2. Quantifying model risk

2.1. Scoring functions for volatility models

Forecast performances of competing models can be evaluated based on scoring functions. A scoring (loss) function is defined as a function $S : \mathbb{R}_+ \times \mathcal{H} \to \mathbb{R}_+$ where $\mathcal{H}$ is a compact subset of $\mathbb{R}_{++}$, and $\mathbb{R}_+$ and $\mathbb{R}_{++}$ represent the non-negative and positive parts of the real line, respectively.

In terms of model comparisons, the evaluation of volatility forecasting models depends on the choice of variance proxy $\hat{\sigma}^2$ and of the scoring function S. To contrast two sets of competing variance forecasts, $\{h^k\}$ and $\{h^j\}$, of model k and j, over a period from t to $t+\tau$, we compute and compare the expected scores of the two models: $\mathbb{E}\left[S(\hat{\sigma}^2, h^k)\right] = \frac{1}{\tau+1} \cdot \sum_{i=t}^{t+\tau} S(\hat{\sigma}_i^2, h_i^k)$ and $\mathbb{E}\left[S(\hat{\sigma}^2, h^j)\right] = \frac{1}{\tau+1} \cdot \sum_{i=t}^{t+\tau} S(\hat{\sigma}_i^2, h_i^j)$, given a variance proxy $\hat{\sigma}^2$ and a scoring function S. Throughout this paper, we use the population estimate for average values. A smaller expected score indicates a superior forecasting ability of the volatility model. Also, for a given scoring function and variance proxy, the optimal variance forecast denoted by h_t^* can be obtained by minimizing the expected score and is defined below, where $\mathcal{F}_{t-1}$ denotes the information set at $t-1$ (see Patton, 2011; this is further generalized for point forecasts in Gneiting, 2011):

$$h_t^* \equiv \underset{h \in \mathcal{H}}{\arg\min}\, \mathbb{E}\left[S(\hat{\sigma}_t^2, h) | \mathcal{F}_{t-1}\right]. \tag{1}$$

We consider the MSE and QLIKE scoring functions (denoted by S_{mse} or S_{qlike}) in this paper. The robustness property of scoring functions distinguishes these two scoring functions in that the ordering of any two (possibly imperfect) volatility forecasts by the expected score of MSE or QLIKE is the same whether the ordering is done using the true conditional variance or some conditionally unbiased variance proxy (Patton, 2011). These two prominent robust scoring functions are listed below, when a conditionally unbiased variance proxy $\hat{\sigma}^2$ is used:

$$\text{MSE} : S_{mse}(\hat{\sigma}^2, h) = (\hat{\sigma}^2 - h)^2; \quad \text{QLIKE} : S_{qlike}(\hat{\sigma}^2, h) = \log(h) + \frac{\hat{\sigma}^2}{h} \quad (2)$$

2.2. Measuring model risk of volatility models

In the following, we quantify the model risk of a volatility model j provided that a time series of out-of-sample daily variance forecasts $h_t^j, ..., h_{t+T}^j$ (computed in our case using rolling windows) as well as a time series of observed daily variance proxy $\hat{\sigma}_t^2, ..., \hat{\sigma}_{t+T}^2$ for time t, $t+1$,..., $t+T$ are used. We start with a definition of true model risk of volatility models. In practice, this cannot be used because the true variance process is unknown. We then propose a model risk proxy measure using a conditionally unbiased variance proxy. Then we proceed with proposing measures of model risk of volatility models, differing via their structure (additive vs. multiplicative). We discuss how these measures depend on subjective inputs (optimization window length, model risk evaluation window length, variance proxy as well as the loss function).

Definition 1. If the sequence of true variances $\{\sigma^2\}$ is known, and the volatility forecaster produces a time series of conditional variance forecasts $\{h^j\}$ with volatility model j, then the *true model risk* of model j over a model risk evaluation window of length $n+1$ from t to $t+n$ is quantified by $p_{[t,t+n]}^j$:

$$p_{[t,t+n]}^j = \frac{1}{n+1} \cdot \sum_{i=t}^{t+n} \left| \sigma_i^2 - h_i^j \right|. \quad (3)$$

This measure is derived based on absolute differences, whereas an alternative measure based on squared differences is considered in Section 6.1.

In practice, the true variance σ^2 is unobservable but can be replaced by the observed variance proxy $\hat{\sigma}^2$ such as the squared return or the realized variance. Thus, $\hat{p}_{[t,t+n]}^{proxy,j}$ denotes the *model risk proxy* for true model risk $p_{[t,t+n]}^j$ of model j, using a conditionally unbiased variance proxy $\hat{\sigma}^2$:

$$\hat{p}_{[t,t+n]}^{proxy,j} = \frac{1}{n+1} \cdot \sum_{i=t}^{t+n} \left| \hat{\sigma}_i^2 - h_i^j \right|. \quad (4)$$

Although the model risk proxy $\hat{p}^{proxy,j}_{[t,t+n]}$ can be readily computed, being the simplest estimate of model risk, we will show that our proposed model risk estimates based on scoring functions prove to be superior to this model risk proxy. The novelty of our proposed measure is that it uses scoring functions and we consider two different formulations of objective functions. Thus, we propose (i) an additive structure and (ii) a multiplicative structure as discussed below:

(i) *Additive structure:* given a volatility model j, based on (5) we find optimized constants $c^{add,j}_{S,t+\tau+m}$ (which, when added to a series of variance forecasts $\{h^j_i\}^{t+\tau+m}_{i=t+m}$, minimize the expected score of a scoring function S over an optimization window from $t+m$ to $t+\tau+m$ of length $\tau+1$), where $m = 0 : T-\tau$. The constant c^{add} is restricted so that $h^j_i + c^{add} > 0$ is satisfied for all i in order to ensure the positivity of variance forecasts:

$$c^{add,j}_{S,t+\tau+m} = \underset{c^{add}}{\arg\min} \frac{1}{\tau+1} \cdot \sum_{i=t+m}^{t+\tau+m} S\left(\hat{\sigma}^2_i, h^j_i + c^{add}\right). \tag{5}$$

As the optimization window of length $\tau+1$ is rolled forward at every step, a time series of optimized increments $\{c^{add,j}_{S,i}\}^{t+T}_{i=t+\tau}$ is generated for the variance forecasts of model j. Subsequently, the *model risk estimate* of model j over a model risk evaluation window from $t+\tau$ to $t+\tau+n$ under an additive structure is given by $\hat{p}^{add,j}_{S,[t+\tau,t+\tau+n]}$:

$$\hat{p}^{add,j}_{S,[t+\tau,t+\tau+n]} = \frac{1}{n+1} \cdot \sum_{i=t+\tau}^{t+\tau+n} \left|(h^j_i + c^{add,j}_{S,i}) - h^j_i\right|. \tag{6}$$

(ii) *Multiplicative structure:* we calculate optimized multipliers $c^{mul,j}_{S,t+\tau+m}$ that are assigned to the conditional variance forecasts $\{h^j_i\}^{t+\tau+m}_{i=t+m}$ via minimizing the expected score over an optimization window from $t+m$ to $t+\tau+m$ with window length $\tau+1$, where $m = 0 : T-\tau$ and the constant c^{mul} is constrained to satisfy $c^{mul} > 0$ for the positivity of variance forecasts. This is given below:

$$c^{mul,j}_{S,t+\tau+m} = \underset{c^{mul}}{\arg\min} \frac{1}{\tau+1} \cdot \sum_{i=t+m}^{t+\tau+m} S\left(\hat{\sigma}^2_i, h^j_i \cdot c^{mul}\right). \tag{7}$$

Then under a multiplicative structure the *model risk estimate* of model j is given by $\hat{p}^{mul,j}_{S,[t+\tau,t+\tau+n]}$:

$$\hat{p}^{mul,j}_{S,[t+\tau,t+\tau+n]} = \frac{1}{n+1} \cdot \sum_{i=t+\tau}^{t+\tau+n} \left| (h^j_i \cdot c^{mul,j}_{S,i}) - h^j_i \right|. \tag{8}$$

In this framework, model risk is computed as the average model error in terms of volatility forecasting, with the size of model error obtained via scoring function minimizations. If these minimizations indicate that model error is small (the volatility forecasts requiring only a small adjustment to minimize the score), then the volatility model has a low model risk.

A similarity can be drawn between equations (5) and (7) and the Mincer-Zarnowitz regressions for variance forecast tests (Hansen and Lunde, 2005). For these, the variance proxy is regressed on the variance forecast, and the optimality of the forecasts is verified via the parameter values (0 for the intercept and 1 for the sensitivity). Similarly, in our case one would expect c^{add} to be 0 and c^{mul} to be 1, which would lead to zero model risk. For brevity, we will omit the subscripts for time intervals for $p^j_{[t+\tau,t+\tau+n]}$, $\hat{p}^{proxy,j}_{[t+\tau,t+\tau+n]}$, $\hat{p}^{add,j}_{S,[t+\tau,t+\tau+n]}$ and $\hat{p}^{mul,j}_{S,[t+\tau,t+\tau+n]}$. In order to detect the similarity of the model risk proxy (4) and model risk estimates (6) and (8) to the true model risk measure (3), we compute several measures of similarity and discuss the results in the following simulation study.

The computation of model risk depends on five subjective choices regarding: 1) the optimization framework (additive or multiplicative), 2) the optimization window length τ, 3) the model risk evaluation window length n, 4) the loss function (MSE or QLIKE) and 5) the variance proxy. In the following, we study each of these alternatives via Monte Carlo simulations and identify the optimal way to estimate model risk.

3. Simulation study

In this section, we verify via simulations how closely the model risk estimates are able to capture the size of true model risk for various volatility models. Considering that the conditional distribution of financial time series is often fat-tailed and asymmetric, we use the GARCH(1,1) model with skewed Student's t distributed innovations (SKTGARCH), allowing for kurtosis and skewness, with the data generating process specified as:

$$\begin{aligned} r_t &= \sqrt{h_t} Z_t, \quad Z \sim \text{skewed Student's } t\ (\nu, \lambda), \\ h_t &= \hat{\omega} + \hat{\alpha} r^2_{t-1} + \hat{\beta} h_{t-1}, \end{aligned} \tag{9}$$

where r_t denotes a realization of return and h_t denotes the one-step ahead conditional variance forecast for time t. The density function of the standardized returns Z is $f(z|\nu, \lambda)$ (see Appendix Appendix A), in which ν is the degrees of freedom parameter and λ is the skewness parameter. The model parameter estimates are constrained to satisfy the following conditions $\hat{\omega}, \hat{\alpha}, \hat{\beta} > 0, \hat{\alpha} + \hat{\beta} < 1, 2 < \hat{\nu} < \infty$ and $-1 < \hat{\lambda} < 1$. We obtain model parameters by estimating the model on the S&P500 index daily returns from 2000/01/03 to 2010/12/31 (2869 observations): $\hat{\omega} = 7.8183e^{-07}, \hat{\alpha} = 0.0770, \hat{\beta} = 0.9205, \hat{\nu} = 7.1845$ and $\hat{\lambda} = -0.0848$.

Table 1: Volatility models for daily conditional variance forecasts h_t.

RW250:	$h_t = \frac{1}{249} \sum_{i=t-250}^{t-1} \left(r_i - \frac{1}{250} \sum_{i=t-250}^{t-1} r_i\right)^2$
RW1000:	$h_t = \frac{1}{999} \sum_{i=t-1000}^{t-1} \left(r_i - \frac{1}{1000} \sum_{i=t-1000}^{t-1} r_i\right)^2$
RiskMetrics:	$h_t = (1-\lambda) r_{t-1}^2 + \lambda h_{t-1}$, where $\lambda = 0.94$
ARCH(1):	$h_t = \omega + \alpha r_{t-1}^2$
GACRH(1,1):	$h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1}$
EGARCH(1,1):	$\log(h_t) = \omega + \alpha \left[\frac{\lvert r_{t-1}\rvert}{\sqrt{h_{t-1}}} - \mathbb{E}\{\frac{\lvert r_{t-1}\rvert}{\sqrt{h_{t-1}}}\}\right]$
	$+\kappa(\frac{\lvert r_{t-1}\rvert}{\sqrt{h_{t-1}}}) + \beta \log(h_{t-1})$
GJR-GARCH(1,1):	$h_t = \omega + \alpha r_{t-1}^2 + \xi \mathbb{1}\{r_{t-1} < 0\} r_{t-1}^2 + \beta h_{t-1}$

Note. For all (G)ARCH specifications, the daily return $r_t = \sqrt{h_t} Z_t$. Z_t denotes the standardized return and follows the normal, Student's t, skewed Student's t and generalized error distributions, with details shown in Appendix A.

The steps of the simulation analysis are summarized as follows:

Step 1: Generate a time series of 10000 daily returns using the parameter values above.

Step 2: Using the first window of 1000 simulated returns, we estimate 19 volatility models as specified in Table 1 to make one-step ahead conditional variance forecasts (see Hansen and Lunde (2005) for a comprehensive review of volatility models) by using rolling windows with length 1000 (except for the RW250 method for which we use a window length of 250). More precisely, the models are: 1) historical volatility measures with window length of 250 and 1000 (RW250 and RW1000), which are nonparametric; 2) the RiskMetrics model with $\lambda = 0.94$, denoted by RiskMetrics; 3) the autoregressive conditional heteroscedasticity (ARCH(1)) models (Engle, 1982) with one lag, combined with four specifications for the standardized errors following the normal (N), Student's t (T), skewed Student's t (SKT) and generalized error distributions (GED); and 4) generalized autoregressive conditional heteroscedasticity (GARCH) models combined with the aforementioned four distributional assumptions for the standardized errors, including the symmetric GARCH(1,1) models (Bollerslev, 1986), as well as the EGARCH(1,1) (Nelson, 1991) and GJR-GARCH(1,1) (Glosten et al., 1993) models with leverage terms to consider asymmetry in volatility clustering. The notations for the (G)ARCH specifications are written as the combination of acronyms for distributional assumptions and ARCH, GARCH, EGARCH, or GJR; for example, NARCH stands for the ARCH(1) model with normal innovations.

Step 3: We use the next (up to) 2000 simulated returns as the optimization window, obtaining the additive or multiplicative optimized constants as in equation (5) or (7).

Step 4: Then, using the next (up to) 1000 simulated returns as the model risk evaluation window, we compute the model risk estimates as in equations (6) and (8).

Specifically, for each model we use the squared return as the variance proxy for a given scoring function (S_{mse} or S_{qlike}) and calculate the model risk estimates of daily volatility forecasts, considering several optimization windows of length $\tau_1 = 250$, $\tau_2 = 500$, $\tau_3 = 1000$ and $\tau_4 = 2000$ and two model risk evaluation windows of length $n_1 = 250$ and $n_2 = 1000$.

Step 5: We roll the window one step ahead and repeat steps 2 to 4, until we reach the end of the simulated series.

As an output, we obtain a series of a length of 6,000 (obtained as $10000 - 1000 - 2000 - 1000$, as described in Steps 1 to 4 above) of time-varying model risk estimates, and in the following we study these model risk estimates.

To analyze the degree of similarity of the model risk proxy and our proposed model risk estimates to the true model risk, we use Pearson's linear correlation coefficient $\mathcal{C}^{\mathcal{M}} = Correl(p^{\mathcal{M}}, \hat{p}^{proxy,\mathcal{M}}$ or $\hat{p}_S^{\mathcal{M}})$ between the true model risk ($p^{\mathcal{M}}$) and model risk proxy ($\hat{p}^{proxy,\mathcal{M}}$) as well as model risk estimates ($\hat{p}_S^{\mathcal{M}}$) across the set of volatility models $\mathcal{M}$ presented in Table 1, in which $\hat{p}_S^{\mathcal{M}}$ refers to $\hat{p}_S^{add,\mathcal{M}}$ or $\hat{p}_S^{mul,\mathcal{M}}$ under two different formulations of objective function. Pearson's correlation coefficient can only show a linear relationship between two series, so to consider a possibly nonlinear association between true model risk and model risk proxy or model risk estimates, we compute our second measure of similarity as $\tau_x^{\mathcal{M}} = \tau_x(p^{\mathcal{M}}, \hat{p}^{proxy,\mathcal{M}}$ or $\hat{p}_S^{\mathcal{M}})$ the rank correlation coefficient of Emond and Mason (2002) that extends Kendall's nonparametric measure τ_b. As the third measure of similarity, we consider the explanatory power of the measures of model risk: for a volatility model j , $\psi^j = \hat{p}^{proxy,j}/p^j$ or $\hat{p}_S^j/p^j$, where $\hat{p}_S^j$ can be $\hat{p}_S^{add,j}$ or $\hat{p}_S^{mul,j}$.

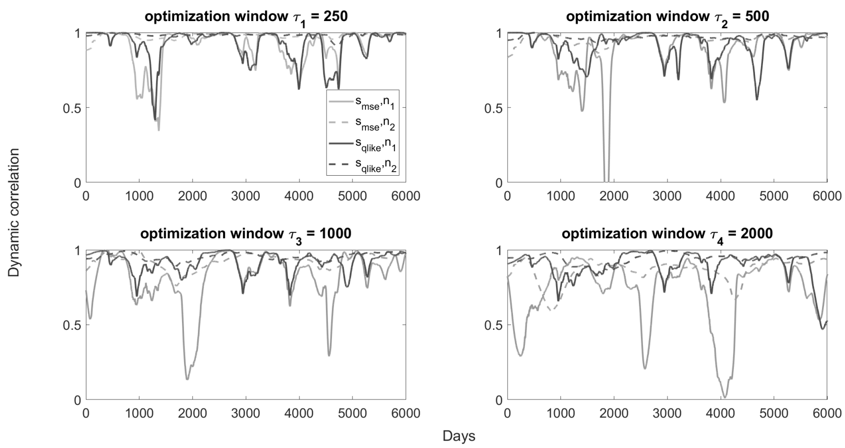
3.1. Optimization framework

This subsection compares the performance of the additive and multiplicative structures for the model risk given in formulae (6) and (8). Based on returns simulated with the SKTGARCH model, panel (a) of Figure 1 presents the dynamic correlation $\mathcal{C}^{\mathcal{M}}$ between the true model risk p in (3) and the model risk estimates $\hat{p}_S^{add}$ in (6) based on the MSE and QLIKE loss functions under an additive structure across all models considered in this paper, whilst panel (b) shows the dynamic correlation between the true model risk p and model risk estimates $\hat{p}_S^{mul}$ in (8) under a multiplicative structure. We can conclude that the model risk estimate based on a multiplicative structure leads to lower correlations between the true and estimated model risk. Thus, in the remaining part of the paper, we estimate model risk using an additive structure only and present the corresponding results, and write this model risk estimate $\hat{p}_S^{add,j}$ as $\hat{p}_S^j$ for brevity, given model j.

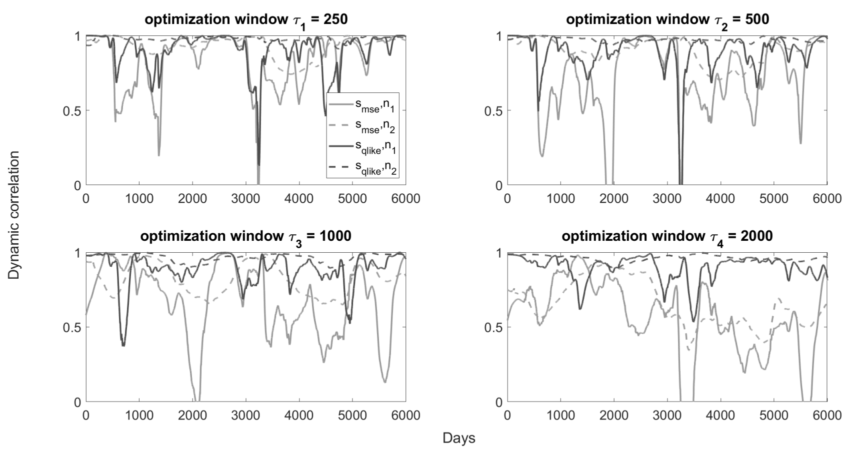
3.2. Optimization window length

This subsection studies the effect of the length of the optimization window on the model risk estimates. Figure 2 juxtaposes the QLIKE-based model risk (dashed lines) with true model risk (solid line). By fixing the distributional assumption to the true data generating process (SKTGARCH), the true model risk is essentially parameter estimation risk which arises due to inaccurate parameter estimation of volatility models. The QLIKE-based model risk is computed over multiple optimization windows ($\tau_2 = 500$ and $\tau_3 = 1000$) and a fixed model risk evaluation window of length 250, and using the squared return proxy. As seen in this figure, the effect of optimization window length is nontrivial in estimating the model risk of variance estimates. Figure 3 shows the quantile plots of the correlations between true model risk and QLIKE-based model risk estimates across the set of volatility models. This highlights that, in 95% of the cases, the estimated model risk has a correlation of 0.75 or higher with the true model risk.

Table 2 reports the average values of Pearson's correlation, the τ_x correlation coefficient and the explanatory power, denoted by $\bar{\mathcal{C}}^{\mathcal{M}}$, $\bar{\tau}_x^{\mathcal{M}}$ and $\bar{\psi}^{\mathcal{M}}$ respectively. In panel A, we report the results for the model risk estimates based on the MSE and QLIKE loss functions, considering the squared return as the variance proxy. In terms of the length of optimization windows, the QLIKE-based model risk estimation methodology using a shorter window can explain a higher proportion of true model risk than that using a longer window, as seen in the last column. Nevertheless, when the optimization window is small (e.g. $\tau_1 = 250$), the QLIKE-based model risk tends



(a) Under an additive structure



(b) Under a multiplicative structure

Figure 1: Dynamic correlation between true model risk p and model risk estimates, $\hat{p}_S^{add}$ or $\hat{p}_S^{mul}$, shown in panel (a) and panel (b) across various volatility models considered in Table 1, based on returns simulated by the SKTGARCH model. The model risk of daily volatility forecasts is computed using the scoring function $S = S_{mse}$ or S_{qlike}, and the squared return is used as the variance proxy. We consider optimization windows of $\tau_1 = 250$, $\tau_2 = 500$, $\tau_3 = 1000$ and $\tau_4 = 2000$ and model risk evaluation windows of $n_1 = 250$ and $n_2 = 1000$.

to overestimate the true model risk. Generally, with the optimization window length increasing, the average Pearson's linear correlation $\bar{\mathcal{C}}^{\mathcal{M}}$ between model risk estimates and true model risk decreases, also seen in Figure 3,

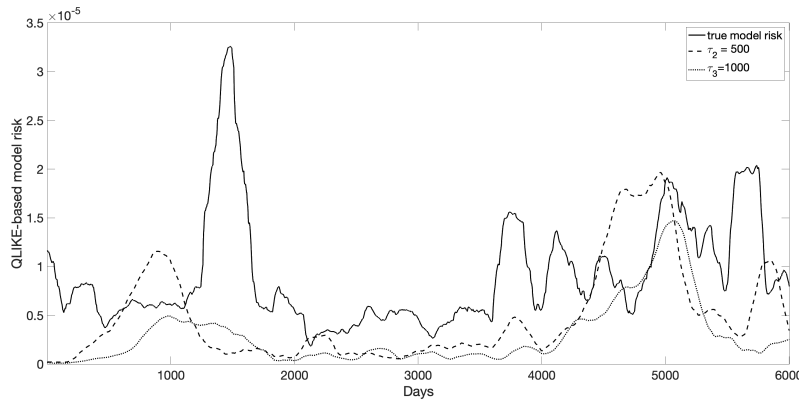


Figure 2: QLIKE-based model risk, calculated over multiple optimization windows (τ_2 and τ_3) and a model risk evaluation window fixed at 250 and using a simulated path of 10000 data points by SKTGARCH.

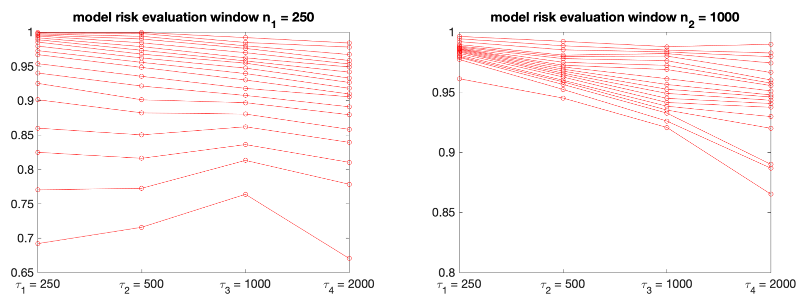


Figure 3: Quantiles of correlations between true model risk and QLIKE-based model risk estimates computed using different lengths of optimization windows (τ_1, τ_2, τ_3, and τ_4) under an additive structure and model risk windows (n_1 and n_2). Each line represents a quantile from 5% to 95% with an increment of 5%.

whilst the average nonlinear rank correlation $\bar{\tau}_x^{\mathcal{M}}$ increases. Based on the correlation results, the shorter optimization window is preferred.

3.3. Model risk evaluation window length

This subsection studies the effect of the length of the model risk evaluation window on model risk estimates. We find that the model risk estimates based on a longer model risk evaluation window ($n_2 = 1000$) have a higher linear (and nonlinear) correlation with the true model risk, as evidenced by Figure 3 and Table 2. Figure 4 illustrates the average percentage of true model risk explained by the QLIKE-based model risk estimate under an additive structure, calculated using an optimization window of length $\tau_2 = 500$

Table 2: Similarity of model risk estimates and proxy to the true model risk.

Panel A: Similarity of the MSE and QLIKE-based model risk estimates, $\hat{p}_{s_{mse}}$ and $\hat{p}_{s_{qlike}}$					
Model risk estimate	optimization window length	model risk window length	$\bar{\mathcal{C}}^{\mathcal{M}}$	$\bar{\tau}_x^{\mathcal{M}}$	$\bar{\psi}^{\mathcal{M}}$
$\hat{p}_{s_{mse}}$	$\tau_1 = 250$	$n_1 = 250$	0.91	0.68	115%
		$n_2 = 1000$	0.96	0.67	88%
	$\tau_2 = 500$	$n_1 = 250$	0.87	0.73	92%
		$n_2 = 1000$	0.94	0.88	60%
	$\tau_3 = 1000$	$n_1 = 250$	0.82	0.79	66%
		$n_2 = 1000$	0.91	0.87	44%
	$\tau_4 = 2000$	$n_1 = 250$	0.73	0.88	41%
		$n_2 = 1000$	0.86	0.94	29%
$\hat{p}_{s_{qlike}}$	$\tau_1 = 250$	$n_1 = 250$	0.89	0.65	96%
		$n_2 = 1000$	0.98	0.61	85%
	$\tau_2 = 500$	$n_1 = 250$	0.88	0.81	65%
		$n_2 = 1000$	0.97	0.95	50%
	$\tau_3 = 1000$	$n_1 = 250$	0.92	0.88	43%
		$n_2 = 1000$	0.96	0.99	34%
	$\tau_4 = 2000$	$n_1 = 250$	0.89	0.92	32%
		$n_2 = 1000$	0.94	1.00	26%
Panel B: Similarity of the model risk proxy, $\hat{p}^{proxy}$					
$\hat{p}^{proxy}$		$n_1 = 250$	0.35	1.00	832%
		$n_2 = 1000$	0.44	1.00	763%

Note. Calculations are based on daily returns simulated by the SKTGARCH model. $\bar{\mathcal{C}}^{\mathcal{M}}$ and $\bar{\tau}_x^{\mathcal{M}}$ represent average values of linear and nonlinear association between the true and estimated model risk; $\bar{\psi}^{\mathcal{M}}$ shows the average explanatory power of model risk estimates across the set of volatility models. We consider different optimization windows (τ_1, τ_2, τ_3 and τ_4) and model risk evaluation windows (n_1 and n_2), and use the squared return as the variance proxy.

and model risk estimation windows of length $n_1 = 250$ and $n_2 = 1000$. Across all the models considered, the model risk estimate computed over a shorter model risk estimation window captures a larger part of true model risk compared with the estimates computed over a longer window. Additionally, our QLIKE-based model risk estimate explains up to about 80% of true model risk.

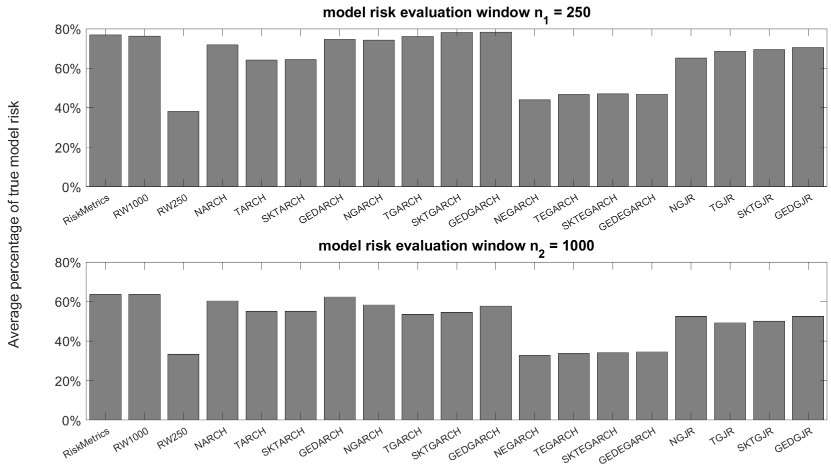


Figure 4: Average percentage of true model risk p explained by the QLIKE-based model risk estimate $\hat{p}_{s_{qlike}}$, computed using an optimization window of length $\tau_2 = 500$.

3.4. Choice of scoring function

This subsection discusses the effect of scoring functions on model risk estimates. In panel A of Table 2, we find that the model risk estimate based on the QLIKE loss function outperforms the one based on the MSE loss function, as the former generally has a higher correlation (and τ_x coefficient) with the true model risk for a given optimization window and model risk evaluation window. The QLIKE-based estimate is highly consistent with the true model risk with a correlation averaging from 0.88 to 0.98.

3.5. Variance proxy

The volatility model risk estimates can be affected by the choice of variance proxy. To investigate this, the squared return and realized variance are considered as alternative proxies. The stochastic volatility model without jumps is used to simulate high-frequency returns for the computation of realized variances, as specified in Huang and Tauchen (2005):

$$\begin{aligned} ds\,(t) &= \mu dt + \exp\left[\beta_0 \;+\; \beta_1 v\,(t)\right] dW_p\,(t)\,, \quad \text{and} \\ dv\,(t) &= \alpha_v v\,(t)\,dt + dW_v\,(t)\,. \end{aligned} \tag{10}$$

$s(t)$ represents the log price process and $v(t)$ is a stochastic volatility factor. The standard Brownian motions W_p and W_v have $correl\,(dW_p, dW_v) = \rho$. Following the setting in Huang and Tauchen (2005), the model parameters are given as: $\mu = 0.03$, $\beta_0 = 0$, $\beta_1 = 0.125$, $\alpha_v = -0.1$, and $\rho = -0.62$. We use the Euler method to simulate 23,400 intervals for each of 10000 days. The simulated realized variances are obtained by aggregating the 5-min squared returns. Based on the simulated data, the model risk of the NARCH model is calculated using the QLIKE loss function under an additive structure, varying the optimization window length in Figure 5, when the model risk evaluation window is fixed at 250. Additionally, we examine the variation of the estimated model risk against the model risk evaluation window length in Figure 6, when the optimization window is fixed at 500. Figures 5 and 6 compare the model risk estimated using different variance proxies with the true model risk and show that, in small samples, the realized variance proxy works better than the squared return proxy in estimating model risk. The effect of variance proxy on model risk dampens when the window length increases.

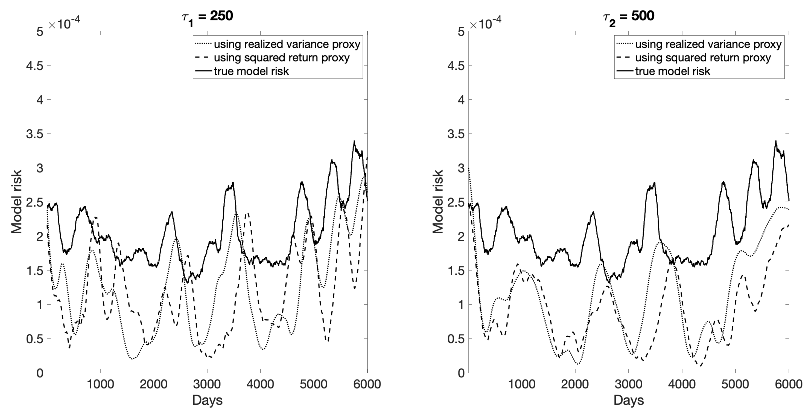


Figure 5: QLIKE-based model risk estimate of the NARCH model compared with the true model risk, computed over multiple optimization windows (τ_1 and τ_2) and a model risk evaluation window of length $n_1 = 250$, based on data simulated using the stochastic volatility model.

The model risk proxy measure, denoted by $\hat{p}^{proxy}$, is computed directly from the distance between the variance proxy (the squared return proxy used) and variance estimate as given in (4). This measure is handy but appears to be flawed as seen in panel B of Table 2. We consider the similarity of the model risk proxy to the true model risk, and find average correlations around 0.35 and 0.44 which are less than half of the corresponding values for the model risk estimates presented in panel A. Also, the model risk proxy tends to overestimate model risk, leading to values that are multiples of

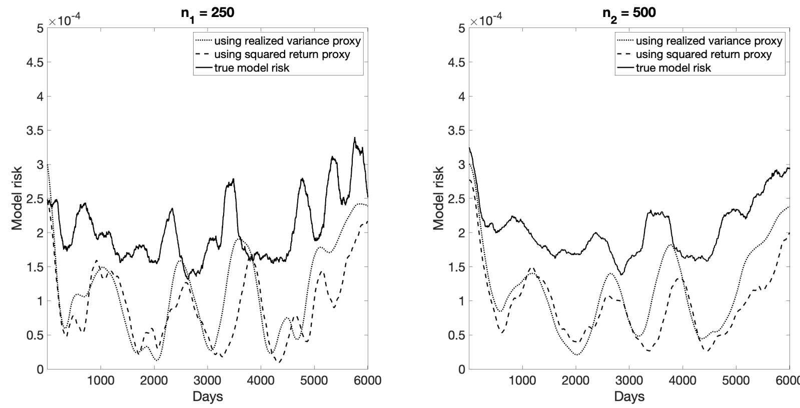


Figure 6: QLIKE-based model risk estimate of the NARCH model compared with the true model risk, computed over multiple model risk evaluation windows (n_1 and n_2) and an optimization window of length $\tau_2 = 500$, based on data simulated using the stochastic volatility model.

the true model risk. From a dynamic perspective, Figure 7 compares the dynamic correlation of the model risk proxy and the QLIKE-based model risk estimate with the true model risk, where the squared return is used as the variance proxy and model risk is computed over an optimization window $\tau_2 = 500$ and a model risk evaluation window $n_1 = 250$. Unlike the QLIKE-based model risk measure, the model risk proxy is unable to give a reliable approximation of true model risk since it is often negatively correlated with the true model risk.

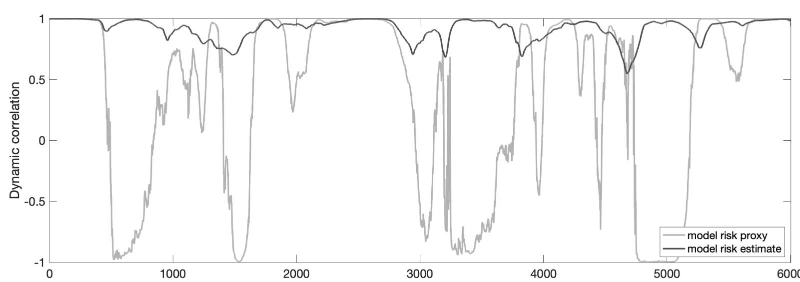


Figure 7: Dynamic correlation of the model risk proxy and the QLIKE-based model risk estimate with the true model risk, computed over a model risk evaluation window of length $n_1 = 250$, based on daily returns simulated by the SKTGARCH model. The squared return is used as the variance proxy.

It is common practice to use the R^2 of the Mincer and Zarnowitz (1969)

(MZ) regression to assess the accuracy (efficiency) of volatility forecasting models, whilst the newly introduced model risk estimates provide a different perspective on the predictive accuracy. It is worthwhile to compare the model risk estimates with the MZ regression results. This regresses the conditionally unbiased proxy ($\hat{\sigma}_t^2$, using the squared returns) on the variance forecast (h_t) of a given model, which is written as $\hat{\sigma}_t^2 = \beta_0 + \beta_1 h_t + e_t$. A higher value of R^2 indicates a better fit of the volatility model to the variance proxy. Figure 8 compares the averages of model risk estimates, true model risk and $(1 - R^2)$ for different volatility models over a simulated path of SKTGARCH returns. The level of model risk estimates is comparable to the $(1 - R^2)$ values. However, whilst the proposed model risk measure successfully identifies the worst performing model RW1000, the MZ method fails to do so. Also, in addition to providing a ranking of competing models, our proposed model risk measure provides actual model risk estimates and it is designed to improve on variance estimates. Furthermore, our measure offers a decomposition of model risk estimates according to the sources of model risk as discussed in Section 5.2.

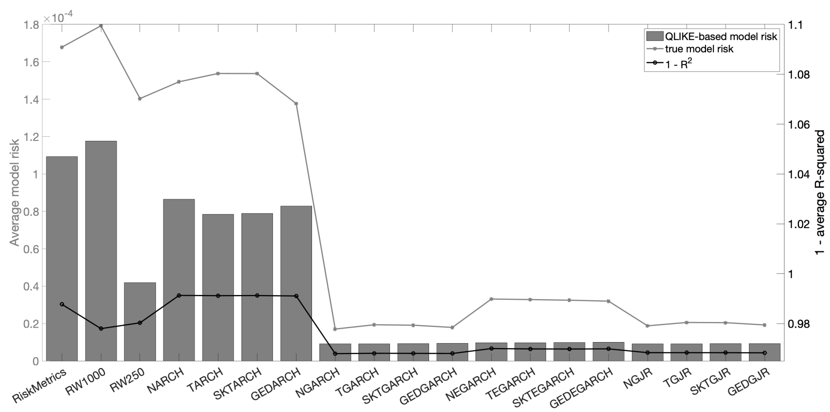


Figure 8: Average QLIKE-based model risk estimated in an optimization window of length $\tau_2 = 500$ and a model risk window of length $n_1 = 250$, compared with $(1-R^2)$, for different volatility models, using data simulated by the SKTGARCH model.

Overall, based on our simulation analysis we find that the scoring function-based model risk estimation methodology that uses the additive structure defined in (6) can be an efficient tool in approximating the true model risk of volatility models. Our results imply that shorter optimization and model risk evaluation windows are preferred. As such, the optimization windows of length $\tau_2 = 500$ or $\tau_3 = 1000$ and a model risk evaluation window of length

$n_1 = 250$ are chosen. Also, the QLIKE scoring function used for optimization provides model risk estimates which are highly correlated with the true model risk, with a correlation averaging from 0.88 to 0.98. In small samples, the effect of variance proxy on estimating the model risk is noticeable with the realized variance proxy outperforming squared returns.

4. Properties of model risk estimates

To facilitate model risk management from the regulators' perspective, a reasonable positive measure $\rho(\cdot)$ of risk should satisfy the coherence properties (McNeil et al., 2015): 1) Monotonicity: for returns r_1 and r_2 with $r_1 \leq r_2$, we have that $\rho(r_1) \geq \rho(r_2)$; 2) Positive homogeneity: for any positive number $a \in \mathbb{R}^+$, we have that $\rho(a \cdot r) = a \cdot \rho(r)$ where r denotes the returns; 3) Translation invariance: for any $a \in \mathbb{R}$, $\rho(r + a) = \rho(r) - a$; and 4) Subadditivity: for any returns r_1 and r_2, $\rho(r_1 + r_2) \leq \rho(r_1) + \rho(r_2)$.

In a similar vein, we focus on the properties of the QLIKE-based model risk measure, denoted by $\hat{p}^j_{S_{qlike}}(r, h^j)$, that assumes an additive structure and uses the squared returns r^2 as the variance proxy. r denotes the daily returns of an asset and h^j denotes the one-step ahead variance forecasts of a model j. Consider the following properties that a reasonable measure of model risk of volatility models should satisfy:

i) *Monotonicity*: If $\sigma^2 < h^i < h^j$ or $\sigma^2 > h^i > h^j$ for all t, then $\hat{p}^i_{S_{qlike}}(r, h^i) < \hat{p}^j_{S_{qlike}}(r, h^j)$, assuming that two different volatility models i and j produce variance forecasts h^i and h^j respectively.
This property states that if the variance forecasts of a certain model are closer to the true variances σ^2, then this model will carry a lower level of model risk.

ii) *Positive homogeneity*: For $a \in \mathbb{R}^+$ and a model j, $\hat{p}^j_{S_{qlike}}(a \cdot r, a^2 \cdot h^j) = a^2 \cdot \hat{p}^j_{S_{qlike}}(r, h^j)$, given variance forecasts h^j of volatility model j.
This states that if the return data is rescaled by a positive constant a and correspondingly the variance forecasts are rescaled by a^2, then the model risk will be resized by a^2 as well.

iii) *Translation invariance*: For a model j, if a constant a with $a > \max_t(\sigma^2_t - h^j_t)$ conditioning on $h^j_t > \sigma^2_t$ for all t, or with $-\min_t(h^j_t) < a < \min_t(\sigma^2_t - h^j_t)$ conditioning on $h^j_t < \sigma^2_t$ for all t, $\hat{p}^j_{S_{qlike}}(r, h^j + a) = \hat{p}^j_{S_{qlike}}(r, h^j) + a \cdot \mathrm{sgn}(h^j - \sigma^2)$.

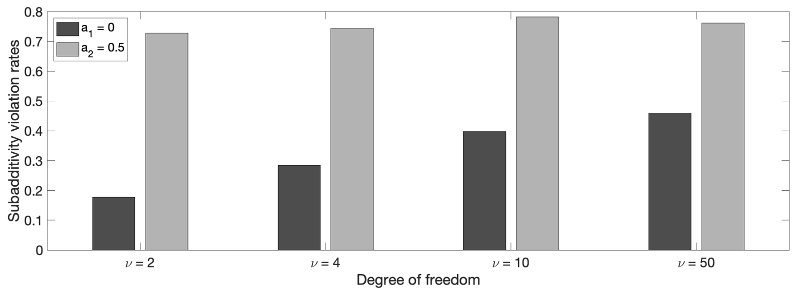


Figure 9: Subadditivity violation rates for RW1000 that produces daily variance forecasts based on simulated returns of assets X and Y and an equally weighted portfolio $(X+Y)$. Assuming that assets X and Z are independent but follow the same Student's t distribution with degrees of freedom ν, asset Y is defined as $Y = aX + \sqrt{1-a^2}Z$ with $a = a_1$ or a_2. The QLIKE-based model risk is computed over an optimization window of $\tau_2 = 500$ and a model risk evaluation window of $n_1 = 250$, using the squared returns as the variance proxy.

This property says that when the variance forecasts are shifted by a constant a that satisfies the conditions above, then the model risk of model j will change with the value of $a \cdot \mathrm{sgn}(h^j - \sigma^2)$.

iv) *Subadditivity*: $\hat{p}^j_{S_{qlike}}(r_{(X+Y)}, h^j_{(X+Y)}) < \hat{p}^j_{S_{qlike}}(r_X, h^j_X) + \hat{p}^j_{S_{qlike}}(r_Y, h^j_Y)$, considering that a model j produces the variance forecasts h^j_X, h^j_Y and $h^j_{(X+Y)}$ when applied to individual assets X and Y, and an equally weighted portfolio $(X+Y)$ consisting of these two assets.
This states that for a given volatility model, the model risk for an equally weighted portfolio comprised of assets X and Y is lower than the sum of model risk for the constituent assets. This property should not be required for measures of model risk of volatility models, as it does not follow the expected behavior of model risk measures.

These properties are difficult to prove theoretically because the estimated model risk depends on the variance proxy used. However, via Monte Carlo simulations, we find that the properties of monotonicity, positive homogeneity and translation invariance hold for the QLIKE-based model risk estimated using the squared return as the variance proxy, whilst, as expected, the subadditivity property does not hold. It must be noted that the properties which hold in our simulation settings are not guaranteed to hold for other data generating processes.

In Figure 9, we revisit the subadditivity of our proposed model risk measure in simulated cases as in Daníelsson et al. (2013), and report the subadditivity violation rates. Specifically, assuming that assets X and Z are independent but follow the same Student's t distribution with degrees of freedom $\nu = 2, 4, 10$, and 50, we consider asset Y correlated with asset X with correlation coefficient a, given by $Y = aX + \sqrt{1-a^2}Z$. We consider two cases: in the first one X and Y are independent ($a_1 = 0$); in the second case X and Y are correlated ($a_2 = 0.5$). We simulate 500 paths of 1750 returns for X and Y, and build an equally weighted portfolio $(X + Y)$. Subsequently, we make one-step ahead variance forecasts of the RW1000 model and compute the QLIKE-based model risk estimates over an optimization window of length $\tau_2 = 500$ and a model risk evaluation window of length $n_1 = 250$ for the individual assets and the portfolio. If the model risk estimate of the portfolio is larger than the sum of individual model risk estimates of the component assets, the subadditivity property is violated for this simulated path. As the results show, the subadditivity violations are very high as expected.

5. Empirical application

Since our model risk estimation framework for volatility models is proved to be efficient via Monte Carlo Simulations, it is of interest to perform an empirical analysis to measure model risk and further dissect model risk into two major sources of risk: model misspecification risk and parameter estimation risk. In this section, we estimate the QLIKE-based model risk under the additive structure using multiple optimization windows. A shorter model risk evaluation period with $n_1 = 250$ is used, since it is in line with the backtesting period of market risk models (Basel Committee on Banking Supervision, 2019), and based on this, the proposed model risk estimate captures a higher proportion of true model risk than using a longer evaluation period of length $n_2 = 1000$.

5.1. Measuring model risk across various assets

We apply the QLIKE-based model risk estimation methodology for several asset classes with daily data (30/12/1983 - 21/10/2019), downloaded from DataStream: 1) FTSE100 index close prices (FTSE100); 2) JP Morgan Chase close prices (JPM); 3) Europe Brent spot prices (dollars per barrel) for Crude Oil (Crude Oil); and 4) Foreign exchange USD/GBP rates (USD/GBP). To consider an alternative variance proxy for the conditionally unbiased variance estimator, we also download the 5-min realized variances of the FTSE100 index (04/01/2000 to 10/10/2019) from the realized library of Oxford-Man Institute of Quantitative Finance, for which Li and

Xiu (2016) provide an empirical justification compared with alternative realized variance estimators. We compute daily log-returns of different assets and then produce out-of-sample daily variance forecasts in a rolling window scheme of length 1000 (except estimates for RW250), given the set of models detailed in Table 1.

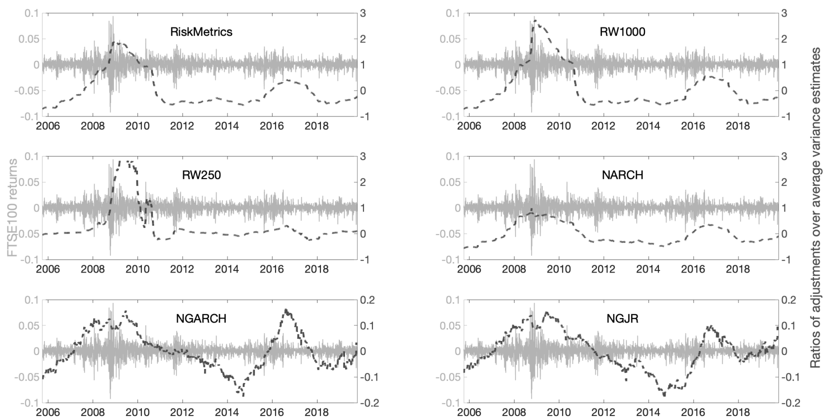
Table 3: Average ratios of the QLIKE-based model risk estimates, based on different variance proxies (squared returns and 5-min realized variances), to estimated variances.

	The volatility proxy $\hat{\sigma}^2$ is							
	squared returns				5-min realized variances			
Models	$\tau_1 = 250$	$\tau_2 = 500$	$\tau_3 = 1000$	$\tau_4 = 2000$	$\tau_1 = 250$	$\tau_2 = 500$	$\tau_3 = 1000$	$\tau_4 = 2000$
RiskMetrics	49.5%	43.6%	31.5%	27.0%	46.9%	42.2%	36.6%	20.9%
RW1000	52.8%	46.3%	30.7%	32.2%	48.9%	43.3%	34.7%	30.1%
RW250	28.8%	14.8%	8.8%	15.6%	25.4%	14.0%	4.6%	7.4%
NARCH	39.3%	35.1%	23.8%	24.2%	39.3%	35.6%	28.2%	17.7%
TARCH	39.7%	35.9%	30.8%	22.8%	40.1%	36.8%	31.9%	16.9%
SKTARCH	39.8%	36.0%	31.1%	22.4%	40.2%	36.9%	32.1%	16.1%
GEDARCH	38.5%	34.5%	24.4%	25.7%	39.5%	35.7%	27.8%	19.6%
NGARCH	11.0%	8.7%	4.2%	2.7%	14.1%	12.2%	10.9%	7.3%
TGARCH	9.5%	7.9%	4.5%	2.8%	13.7%	12.1%	11.2%	7.3%
SKTGARCH	9.2%	7.7%	4.4%	2.4%	13.3%	11.8%	10.9%	7.4%
GEDGARCH	10.1%	8.3%	4.7%	2.6%	13.8%	12.1%	10.9%	7.2%
NEGARCH	10.5%	5.8%	2.5%	4.3%	12.1%	8.0%	2.7%	2.6%
TEGARCH	10.7%	6.6%	2.6%	5.1%	11.6%	7.9%	2.9%	2.4%
SKTEGARCH	10.8%	6.7%	2.8%	5.3%	11.6%	8.0%	3.1%	2.4%
GEDEGARCH	10.3%	5.8%	2.2%	4.4%	11.7%	7.9%	2.6%	2.5%
NGJR	10.1%	8.0%	6.0%	4.3%	13.3%	10.4%	7.0%	3.5%
TGJR	8.8%	7.1%	5.4%	4.1%	12.3%	9.7%	6.4%	3.3%
SKTGJR	8.9%	7.2%	5.4%	3.7%	12.2%	9.6%	6.4%	3.2%
GEDGJR	9.4%	7.5%	5.7%	4.1%	12.8%	10.0%	6.6%	3.3%

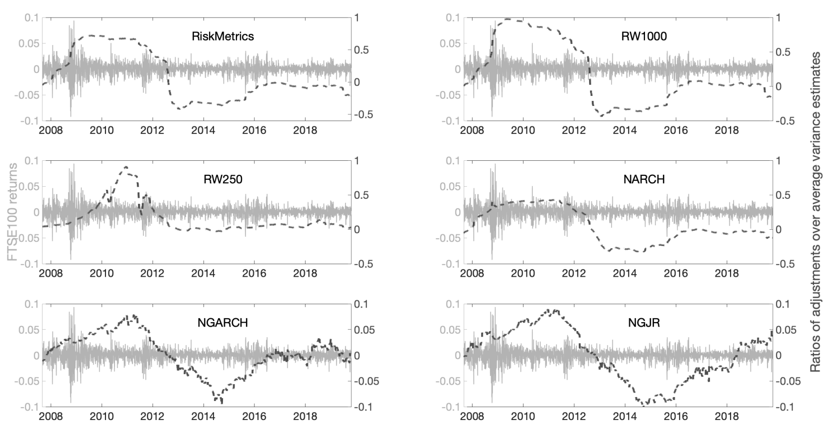
Note. The daily prices and the 5-min realized variances of the FTSE100 index range from 04/01/2000 to 10/10/2019. The optimization window length is $\tau_1 = 250$, $\tau_2 = 500$, $\tau_3 = 1000$ and $\tau_4 = 2000$; the model risk evaluation window length is $n_1 = 250$.

Table 3 reports average ratios of the QLIKE-based model risk estimates computed based on two variance proxies, namely the squared returns and 5-min realized variances, to estimated variances. The purpose of computing the ratio of model risk to the variance forecast of a given model is to make an easy comparison across various volatility models and assets. The model risk estimates are calculated using daily returns and 5-min realized variances of the FTSE100 index from 04/01/2000 to 10/10/2019. When the proposed QLIKE-based model risk is estimated over small optimization windows, the average ratio of model risk over variances, estimated using the realized variance proxy, is in general higher than that using the squared return proxy.

Regardless of the variance proxy used for the computation of the QLIKE-



(a) Optimization window $\tau_2 = 500$



(b) Optimization window $\tau_2 = 1000$

Figure 10: Dynamic additive adjustments made to variance estimates of selected models, obtained based on the QLIKE loss function, for the FTSE100 index returns from 04/01/2000 to 10/10/2019. The variance proxy is the squared return.

based model risk estimates, it is interesting to notice in Table 3 that the model risk estimation method based on a shorter optimization window generally gives higher ratios of model risk estimates than the method based on a longer optimization window. To get a better understanding of this phenomenon, in Figure 10 we compare the additive adjustments with respect to the optimization windows $\tau_2 = 500$ and $\tau_3 = 1000$, and show the time series of adjustments, obtained based on the QLIKE loss function and the squared return used as the variance proxy, made to volatility estimates of selected models. Clearly, the QLIKE-based model risk estimate computed with $\tau_2 = 500$ in panel (a) responds to market events in a more timely and effective manner and allows for a higher level of additive adjustments, which also supports its higher explanatory power in the simulation study, as compared with the estimate computed over $\tau_3 = 1000$ days presented in panel (b) of Figure 10. Therefore, in terms of the QLIKE-based model risk estimates, an optimization window of length $\tau_2 = 500$ is recommended to warrant effective adjustments for model risk and high consistency with true model risk.

Figure 11 shows the time-varying ratios of the QLIKE-based model risk estimates of various models to variance forecasts where the variance proxy used is the squared return. Model risk is estimated over $n_1 = 250$ trading days with an optimization window of length $\tau_2 = 500$, for FTSE100. Within the sample period, the RiskMetrics method, RW1000 and the ARCH(1)-type models are characterised by higher ratios of model risk over the variance forecasts, compared with the rest of the models considered. Noticeably, when the market is highly volatile, the model risk of volatility models generally increases. For example, the FTSE100 index experiences its most uncertain period around 2009, following which the ratios of estimated model risk to estimated variances reach their peak level around 2010 due to the length of the evaluation period for model risk $n_1 = 250$ (one year).

In the Mincer and Zarnowitz (1969) regression ($\hat{\sigma}_t^2 = \beta_0 + \beta_1 h_t + e_t$), the intercept β_0 and the slope β_1 are estimated (alternatively, see other similar regressions using transformations of latent variables, discussed by Jorion, 1995, Bollerslev and Wright, 2001, and Hansen and Lunde, 2006). The null hypothesis of the forecast optimality is that $H^0 : \beta_0 = 0$ and $\beta_1 = 1$. The R^2 of the regression equation is considered as a criterion for the accuracy (efficiency) of the volatility forecasting model. A higher value of R^2 indicates a better forecasting accuracy of the volatility model.

We use the 5-min realized variance as the endogenous variable, and the variance forecast adjusted for model risk (computed using different variance proxies) as the explanatory variable in the MZ regression. In order to analyze the performance of model risk adjustments, Figure 12 presents the change in the R^2 of the MZ regressions (displayed with bars) when the forecasted variance is replaced by the model risk-adjusted variance forecast in the regression. The change in the R^2 shows a very similar pattern to the average ratio (displayed with lines) of model risk estimate over estimated variance. The

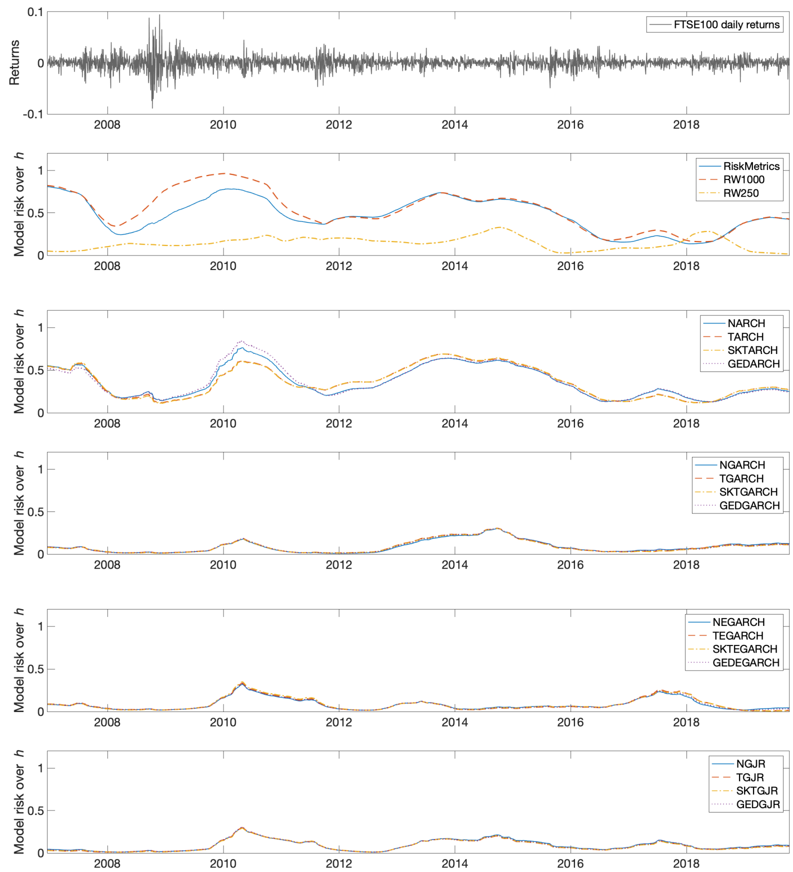
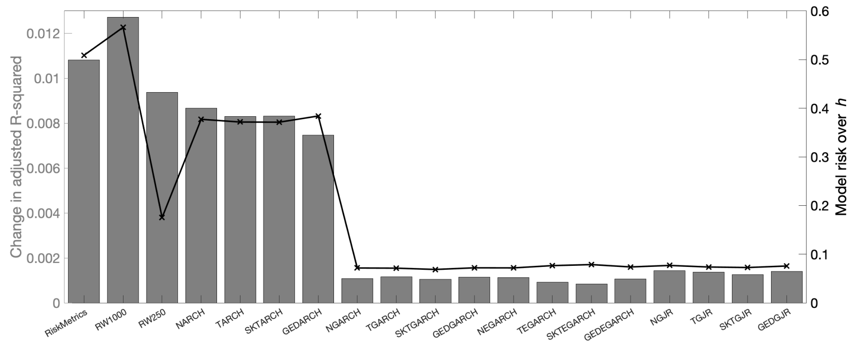
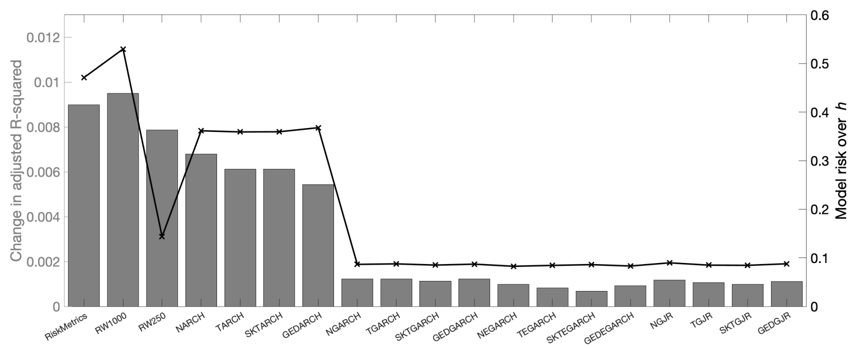


Figure 11: Time-varying ratios of the QLIKE-based model risk estimates to estimated variances, based on the FTSE100 index from 04/01/2000 to 10/10/2019. The squared return is used as the variance proxy. Model risk is computed over $n_1 = 250$ trading days using an optimization window of length $\tau_2 = 500$.

model risk estimates are computed over an optimization window of length $\tau_2 = 500$ and a model risk evaluation window of length $n_1 = 250$, based on the QLIKE loss function, and using the squared return or the realized variance as variance proxy, for the FTSE100 index returns from 04/01/2000 to 10/10/2019. After taking model risk into account, the volatility models have an improved predictive ability as evidenced by an increase in the R^2 across the set of models considered. In general, the higher the model risk, the larger the improvement in the R^2 of the MZ regressions after adjusting for model risk.



(a) Using the squared return as the variance proxy



(b) Using the 5-min realized variance as the variance proxy

Figure 12: Change in the R^2 of the MZ regressions shown in bars after adjusting the variance forecasts for the QLIKE-based model risk of different volatility models using different variance proxies. This is compared to the average ratio of model risk over variance forecasts h, displayed with lines, based on the FTSE index data from 04/01/2000 to 10/10/2019. An optimization window of length $\tau_2 = 500$ and a model risk estimation window of length $n_1 = 250$ are considered.

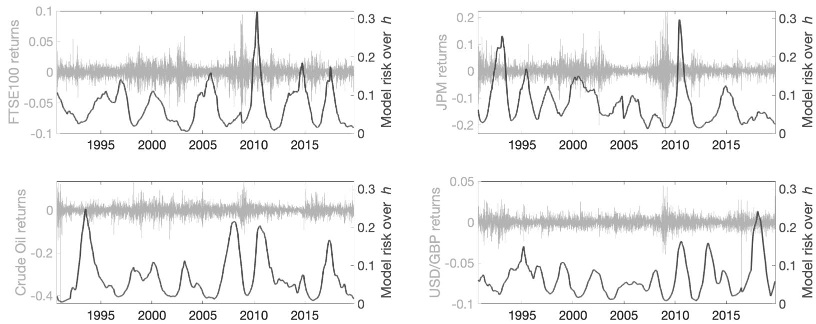
Results for a second application based on several asset classes from 30/12/1983 to 21/10/2019 are shown in Table 4 which presents average ratios of the QLIKE-based model risk estimates to variance forecasts for the set of models in Table 1. Here we use squared returns as the variance proxy and compute the model risk estimates based on optimization windows of length $\tau_1 = 250$, $\tau_2 = 500$, $\tau_3 = 1000$ and $\tau_4 = 2000$, and a model risk evaluation window of length $n_1 = 250$. Noticeably, the longer the optimization window, the lower the level of estimated model risk over variance estimates. For all assets considered, the RW1000 method carries the highest level of

Table 4: Average ratios of the QLIKE-based model risk estimates, with squared returns as the variance proxy, to estimated variances of various models for different assets.

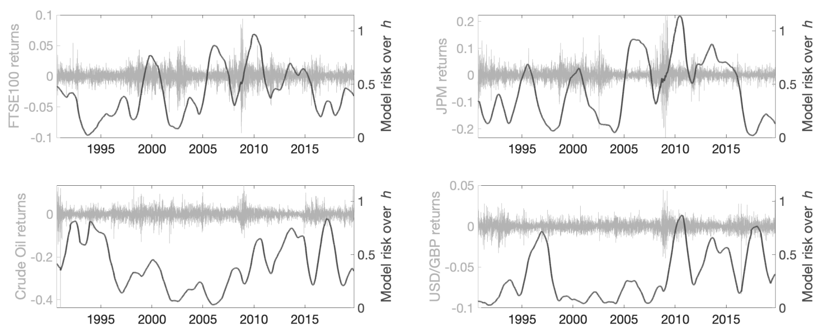
	FTSE100				JPM			
Models	$\tau_1 = 250$	$\tau_2 = 500$	$\tau_3 = 1000$	$\tau_4 = 2000$	$\tau_1 = 250$	$\tau_2 = 500$	$\tau_3 = 1000$	$\tau_4 = 2000$
EWMA	56.8%	46.8%	30.1%	19.2%	59.1%	47.2%	37.0%	22.5%
RW1000	62.7%	51.8%	32.9%	23.2%	67.5%	53.1%	40.8%	31.1%
RW250	27.9%	15.8%	10.4%	9.4%	28.7%	16.9%	11.3%	9.1%
NARCH	43.1%	37.1%	25.3%	17.1%	45.2%	39.7%	34.1%	23.0%
TARCH	42.7%	36.4%	25.9%	15.4%	43.0%	37.4%	32.0%	21.0%
SKTARCH	42.4%	36.1%	25.7%	15.2%	43.1%	37.5%	32.1%	21.0%
GEDARCH	42.9%	37.2%	25.9%	17.6%	43.2%	39.6%	36.0%	28.8%
NGARCH	9.4%	7.6%	5.3%	2.6%	12.4%	8.3%	6.1%	3.2%
TGARCH	9.1%	7.5%	5.4%	2.4%	10.6%	7.0%	5.1%	3.6%
SKTGARCH	8.6%	7.0%	5.1%	2.4%	10.6%	7.0%	5.1%	3.6%
GEDGARCH	9.2%	7.6%	5.5%	2.5%	11.0%	7.2%	5.3%	3.3%
NEGARCH	11.2%	8.1%	4.6%	4.1%	12.3%	9.2%	7.4%	5.2%
TEGARCH	11.4%	8.6%	4.8%	4.5%	11.3%	8.4%	6.7%	6.1%
SKTEGARCH	11.4%	8.6%	4.9%	4.5%	11.3%	8.3%	6.7%	5.9%
GEDEGARCH	11.3%	8.3%	4.6%	4.3%	11.5%	8.6%	7.0%	5.8%
NGJR	9.8%	8.0%	6.0%	3.2%	12.9%	9.9%	7.4%	3.9%
TGJR	9.2%	7.5%	5.6%	2.9%	11.6%	8.5%	6.3%	4.2%
SKTGJR	8.9%	7.3%	5.4%	2.7%	11.6%	8.6%	6.5%	4.1%
GEDGJR	9.6%	7.8%	5.8%	3.1%	11.9%	8.8%	6.5%	3.9%
	Crude oil				USD/GBP			
Models	$\tau_1 = 250$	$\tau_2 = 500$	$\tau_3 = 1000$	$\tau_4 = 2000$	$\tau_1 = 250$	$\tau_2 = 500$	$\tau_3 = 1000$	$\tau_4 = 2000$
EWMA	43.4%	35.1%	22.6%	12.4%	39.0%	32.4%	22.4%	9.5%
RW1000	47.7%	38.7%	25.4%	14.9%	42.7%	35.6%	24.6%	11.6%
RW250	23.7%	15.5%	12.1%	8.6%	23.0%	14.4%	10.9%	7.6%
NARCH	39.9%	32.8%	22.2%	12.2%	33.5%	29.5%	21.6%	9.7%
TARCH	37.9%	30.7%	20.1%	9.8%	36.5%	31.2%	22.6%	11.0%
SKTARCH	38.2%	30.9%	20.2%	10.1%	36.5%	31.3%	22.7%	11.1%
GEDARCH	40.3%	33.0%	22.5%	12.9%	35.7%	31.0%	22.1%	10.3%
NGARCH	9.4%	7.4%	4.7%	3.1%	9.7%	6.9%	4.4%	2.5%
TGARCH	9.1%	6.8%	4.1%	2.9%	9.3%	6.5%	4.3%	2.5%
SKTGARCH	9.2%	6.8%	4.2%	3.1%	9.3%	6.5%	4.2%	2.5%
GEDGARCH	9.3%	7.1%	4.4%	2.9%	9.6%	6.8%	4.6%	2.9%
NEGARCH	13.5%	10.7%	6.7%	4.4%	12.3%	8.9%	5.9%	3.8%
TEGARCH	12.2%	9.3%	5.9%	3.6%	11.8%	8.3%	5.0%	2.9%
SKTEGARCH	12.3%	9.3%	5.9%	3.6%	11.8%	8.2%	4.8%	2.8%
GEDEGARCH	12.9%	10.1%	6.3%	4.0%	11.8%	8.5%	5.5%	3.8%
NGJR	10.7%	8.6%	6.0%	4.7%	11.3%	8.4%	5.3%	3.1%
TGJR	10.3%	7.8%	5.1%	3.9%	10.7%	8.0%	5.0%	2.9%
SKTGJR	10.3%	7.8%	5.1%	3.9%	10.7%	8.0%	4.9%	2.9%
GEDGJR	10.6%	8.2%	5.6%	4.4%	11.0%	8.4%	5.3%	3.4%

Note. Calculations are based on empirical data from 30/12/1983 to 21/10/2019. The model risk is computed over $n_1 = 250$ trading days and multiple optimization windows.

model risk among the set of volatility models, followed by the RiskMetrics method as well as the ARCH(1)-type models. Interestingly, the volatility models have the highest model risk when applied to the JP Morgan Chase stock in general, as compared to the other assets.



(a) SKTGARCH

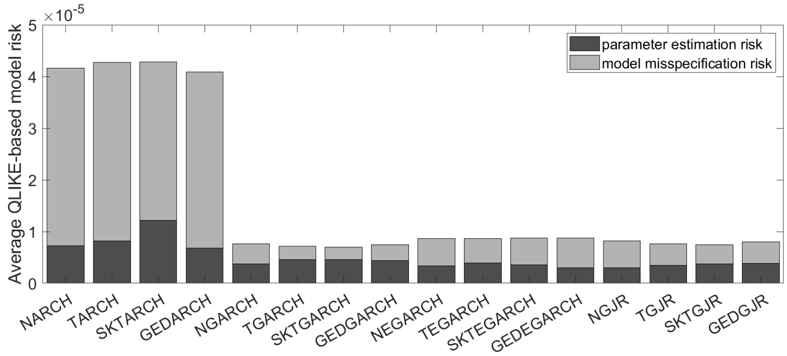


(b) RiskMetrics

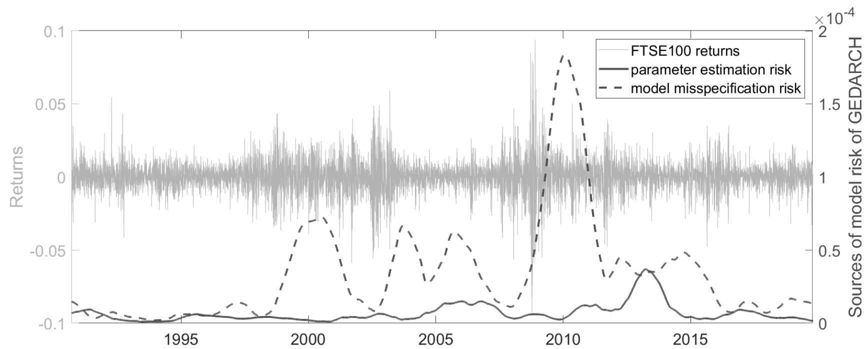
Figure 13: Time-varying ratios of the QLIKE-based model risk estimates to estimated variances, when the SKTGARCH or RiskMetrics model is applied to various assets. The variance proxy is the squared return and model risk is computed over $n_1 = 250$ trading days using an optimization window of length $\tau_2 = 500$, based on data from 30/12/1983 to 21/10/2019.

In Figure 13 we plot the time-varying ratios of the QLIKE-based model risk estimates over variance forecasts of two models: SKTGARCH in panel (a) and RiskMetrics in panel (b), which are applied to various assets from 30/12/1983 to 21/10/2019. We estimate model risk based on a model risk window of length $n_1 = 250$ and an optimization window of length $\tau_2 = 500$, using the squared return as the variance proxy. Comparing the models, SKTGARCH and RiskMetrics, we notice that the ratios of estimated model risk over the variance forecasts fluctuate between 1% and 115% for different assets. Particularly for the returns of equity JP Morgan Chase, the highest ratio of model risk estimates for the RiskMetrics model is around four times higher than for the SKTGARCH model (about 30%). As such, investors need to be aware of the level of model risk of volatility models that particularly increases in uncertain times across various assets.

5.2. Dissecting model risk



(a) Average model risk across various models



(b) Time-varying components of model risk of GEDARCH

Figure 14: Decomposition of the QLIKE-based model risk estimates for various models in panel (a) and for the GEDARCH model in panel (b), based on the FTSE100 returns from 30/12/1983 to 21/10/2019. An optimization window of length $\tau_2 = 500$ and a model risk evaluation window of length $n_1 = 250$ are considered and the variance proxy is the squared return.

The components of model risk are of much interest for the regulatory authorities, practitioners and academics. The major sources of model risk are *parameter estimation risk* and *model misspecification risk* (Kerkhof et al., 2010): parameter estimation risk refers to the uncertainty of parameter estimation; model misspecification risk occurs when the specified model deviates from the true model. Figure 14 disentangles the QLIKE-based model risk estimates of volatility models into these two types of risk for the FTSE100 index from 30/12/1983 to 21/10/2019: panel (a) decomposes the model risk across various models; whilst panel (b) shows the time-varying values of the components of model risk for the GEDARCH model. The calculation of model risk estimates is done over an optimization window of length $\tau_2 = 500$ and a model risk evaluation window of length $n_1 = 250$, and the squared return is used as the variance proxy. For a given volatility model, we compute the parameter estimation risk via simulations of this model with model parameters estimated on the FTSE100 index. Model misspecification risk generally contributes more to the total model risk than parameter estimation risk across various models and over time, as illustrated in panel (a) and panel (b) respectively, though a few exceptions appear when GARCH(1,1) models are considered. When the market becomes volatile, model misspecification risk is aggravated, as seen in panel (b). After the 2008 global financial crisis, the estimate of model misspecification risk noticeably peaks. This empirical evidence calls for the need for model risk management.

In an additional exercise, we investigate the relationship between the constituents of model risk estimates and model-dependent variance forecasts. Based on the daily prices and the 5-min realized variances of the FTSE100 index from 04/01/2000 to 10/10/2019, the QLIKE-based model risk of the set of volatility models is decomposed into model misspecification risk (denoted by $\hat{p}^{MSR}$) and parameter estimation risk (denoted by $\hat{p}^{PER}$). We regress the model misspecification risk estimates $\hat{p}^{MSR}$ and parameter estimation risk estimates $\hat{p}^{PER}$ on explanatory variables related to the following: RET represents the daily return, $RV5$ the 5-min realized variance and Var the variance forecast, all averaged over the previous 250 days; The $Skewness$ and $Kurtosis$ of daily returns are also computed over 250 days. The panel regression equations are written as below:

$$\hat{p}_{it}^{add,MSR} \quad \text{or} \quad \hat{p}_{it}^{add,PER} = X_{it} \cdot \beta + \alpha_i + \epsilon_{it}, \text{ for all } t = 1, ..., T \text{ and model } i \in \mathcal{M}. \tag{11}$$

X_{it} is the time-variant 1×8 regressor vector for all $t = 1$,..., T and it has 8 independent variables shown in panel A including RET, $RV5$, $Skewness \cdot 10^{-4}$, $Kurtosis \cdot 10^{-4}$, $Var - RV5$, $RET \cdot RV5$, $RET \cdot (Var - RV5)$, and $RV5 \cdot (Var - RV5)$. β represents the 8 by 1 vector of coefficients. α_i is the unobserved individual effect for model i, and ϵ_{it} is the error term. $\mathcal{M}$ represents the set of volatility models discussed in Table 1.

The regression results are reported in Table 5; these show an increase of about 18% in the values of adjusted R^2 after adding model specific informa-

Table 5: Panel regression results.

	Model misspecification risk			Parameter estimation risk		
RET	0.013***	0.025***	0.022***	-0.001*	0.000	0.000
	(14.055)	(4.504)	(4.353)	(-1.926)	(0.895)	(0.037)
$RV5$	0.165***	0.159***	0.149***	0.003*	-0.001	-0.002
	(5.003)	(4.527)	(4.868)	(1.885)	(-0.278)	(-1.222)
$Skewness \cdot 10^{-4}$		-0.007***			0.000	
		(-3.142)			(-0.339)	
$Kurtosis \cdot 10^{-4}$		-0.001*			-0.001***	
		(-1.817)			(-3.389)	
$Var - RV5$	0.351***			0.075***		
	(6.445)			(7.236)		
$RET \cdot RV5$	37.845**		8.245***	6.111**		2.004
	(2.330)		(4.381)	(2.798)		(1.237)
$RET \cdot (Var - RV5)$	221.177***			-2.213		
	(3.637)			(-0.388)		
$RV5 \cdot (Var - RV5)$	-447.549			-160.78***		
	(-1.6741)			(-5.331)		
$Adj.R^2$	0.516	0.343	0.339	0.221	0.046	0.025

Note. Panel regression results of equation (10) are based on the daily returns and the 5-min realized variances of the FTSE100 index from 04/01/2000 to 10/10/2019. The model risk of the set of volatility models is computed based on the QLIKE loss function over an optimization window of length $\tau_2 = 500$ and a model risk evaluation window of length $n_1 = 250$, using the squared return as the variance proxy. This table shows the coefficients of model misspecification risk and parameter estimation risk on the variables shown in the first column, associated t-statistics with White (1980)'s standard errors robust to heteroscedasticity adjusted for clusters presented in parentheses and adjusted R^2 reported in the last row of each panel. RET represents the daily return, $RV5$ the 5-min realized variance and Var the variance forecast, averaged over 250 days; the $Skewness$ and $Kurtosis$ of daily returns are computed over 250 days. *, **, *** indicate statistical significance at 10%, 5% and 1% levels, respectively.

tion when explaining model misspecification risk and parameter estimation risk. Model misspecification risk can be explained with an R^2 of about 50%, but parameter estimation risk is more difficult to explain, having an R^2 of less than 25%. To this end, our QLIKE-based model risk estimate provides a measure of the (in)efficiency of volatility models in making volatility forecasts, so provides a practical and reliable measure of model risk.

6. Robustness checks

6.1. Alternative measure of model risk

We consider an alternative definition of model risk measure, i.e. the RMSE formulation based on squared differences instead of the MAE formulation based on absolute differences in Section 2. For example, an alternative definition of model risk is below, rather than using (3):

$$p^j_{[t,t+n]} = \sqrt{\frac{1}{n+1} \cdot \sum_{i=t}^{t+n} \left(\sigma_i^2 - h_i^j\right)^2}. \tag{12}$$

In a similar manner, we can derive RMSE formulations for the model risk proxy and the model risk estimates based on the MSE and QLIKE loss functions to replace (4), (6) and (8).

Table 6: Similarity of the QLIKE-based model risk estimate to the true model risk, using RMSE-based formulations.

Model risk estimate	optimization window length	model risk window length	$\bar{\mathcal{C}}^{\mathcal{M}}$	$\bar{\tau}_x^{\mathcal{M}}$	$\bar{\psi}^{\mathcal{M}}$
$\hat{p}_{s_{qlike}}$	$\tau_1 = 250$	$n_1 = 250$	0.90	0.66	126%
		$n_2 = 1000$	0.97	0.61	115%
	$\tau_2 = 500$	$n_1 = 250$	0.88	0.83	71%
		$n_2 = 1000$	0.95	0.92	56%
	$\tau_3 = 1000$	$n_1 = 250$	0.92	0.92	51%
		$n_2 = 1000$	0.94	1.00	42%
	$\tau_4 = 2000$	$n_1 = 250$	0.89	0.96	44%
		$n_2 = 1000$	0.93	1.00	36%

Note. Calculations are based on data simulated by the SKTGARCH model. The squared return is used as the variance proxy. We consider optimization windows of length τ_1, τ_2, τ_3 and τ_4, and model risk windows of length n_1 and n_2.

Table 6 reports the degree of similarity of the QLIKE-based model risk estimate to the true model risk, in which model risk is computed using RMSE formulations, based on daily returns simulated by the SKTGARCH model. We consider different lengths of optimization windows and model risk windows and use the squared return as the variance proxy. Comparing with Panel A of Table 2, we find that the model risk estimates based on RMSE formulations in Table 6 have similar values of correlations but tend to overestimate the magnitude of model risk, compared with the model risk estimates based on MAE formulations. Thus, our proposed model risk measures based on the MAE formulation are preferable over the alternatives based on an RMSE formulation.

6.2. *Alternative formulation of objective function*

Alternatively, we propose a formulation of objective function combining additive and multiplicative components, based on two constants c^{add} and c^{mul}, as shown below, which combines the additive structure in (5) and the multiplicative structure in (7), for a given volatility model j:

$$(c_{S,t+\tau+m}^{add,j}, c_{S,t+\tau+m}^{mul,j}) = \underset{(c^{add},c^{mul})}{\arg\min} \frac{1}{\tau+1} \cdot \sum_{i=t+m}^{t+\tau+m} S(\hat{\sigma}_i^2, c^{add} + h_i^j \cdot c^{mul}). \quad (13)$$

Conditioning on the non-negativity of variance, we require that $c^{add} + h_i^j \cdot c^{mul} > 0$ with $c^{mul} \geq 0$. Consequently, the corresponding *model risk estimate* of model j is given by $\hat{p}_{S,[t+\tau,t+\tau+n]}^{com,j}$:

$$\hat{p}_{S,[t+\tau,t+\tau+n]}^{com,j} = \frac{1}{n+1} \cdot \sum_{i=t+\tau}^{t+\tau+n} \left| (c_{S,i}^{add,j} + h_i^j \cdot c_{S,i}^{mul,j}) - h_i^j \right|. \quad (14)$$

Table 7: Similarity of model risk estimates computed using an alternative formulation of objective function to the true model risk.

Model risk estimate	optimization window length	model risk window length	$\bar{\mathcal{C}}^{\mathcal{M}}$	$\bar{\tau}_x^{\mathcal{M}}$	$\bar{\psi}^{\mathcal{M}}$
$\hat{p}_{S_{mse}}^{com}$	$\tau_2 = 500$	$n_1 = 250$	0.94	1.00	7446%
		$n_2 = 1000$	0.96	1.00	1155%
	$\tau_3 = 1000$	$n_1 = 250$	0.94	1.00	5224%
		$n_2 = 1000$	0.96	1.00	789%
	$\tau_4 = 2000$	$n_1 = 250$	0.89	1.00	4223%
		$n_2 = 1000$	0.91	0.99	717%
$\hat{p}_{S_{qlike}}^{com}$	$\tau_2 = 500$	$n_1 = 250$	0.92	1.00	9923%
		$n_2 = 1000$	0.94	1.00	1579%
	$\tau_3 = 1000$	$n_1 = 250$	0.91	1.00	5430%
		$n_2 = 1000$	0.93	1.00	844%
	$\tau_4 = 2000$	$n_1 = 250$	0.93	1.00	2372%
		$n_2 = 1000$	0.96	0.99	428%

Note. Calculations are based on data simulated with the SKTGARCH model with parameter values given in Section 3. The squared return is used as the variance proxy. We consider optimization windows τ_2, τ_3 and τ_4, and model risk windows n_1 and n_2.

In Table 7, we present several measures of similarity (the same measures as detailed in Section 3) of model risk estimates, computed using (14), to the true model risk in (3). We find that this alternative model risk measure, though highly correlated with the true model risk, tends to overestimate the true model risk as shown by the explanatory power averaging beyond 100%. This also reenforces us that looking at the correlations solely can be misleading. In this table, the model risk estimate is at least four times higher than the true model risk, leading to too large protection for model risk being set aside compared to how much is needed, when model risk management is performed. This comes with large opportunity costs, which is not favorable. Therefore, this alternative formulation of objective function that combines the additive and multiplicative structure is not recommended for the quantification of model risk of volatility models.

7. Conclusions

In order to assess the accuracy of volatility models, which are of much importance in the financial world, we propose a new model risk measurement framework based on scoring functions, which enables the estimation of model risk of volatility models. This gives a clear indication of the size of model risk of volatility models, directly comparable with the magnitude of the variance estimates given by the models. This is a big advantage because quantifying model risk in this way allows model risk management to be performed.

In a simulation analysis, we consider the effect of using different optimization frameworks, objective functions and alternative variance proxies to compute model risk and compare different lengths of optimization windows and model risk evaluation windows. We recommend the QLIKE-based model risk estimate under an additive structure as a practical and effective measure of true model risk, as we find that this model risk estimate leads to high correlations, averaging from 0.88 to 0.98, between the estimated and true model risk. Particularly the QLIKE-based estimate based on an optimization window of length $\tau_2 = 500$ and a model risk evaluation window of length $n_1 = 250$ is highly consistent with the true model risk, and can explain around 65% of the true model risk across the models. We examine the desirable properties of a reasonable measure of model risk, showing that for the measure of QLIKE-based model risk the required properties are satisfied.

In an empirical study, we explore the effect of different variance proxies on the proposed QLIKE-based model risk estimate, concluding that all else being equal, the level of the the QLIKE-based model risk using the realized variance proxy is generally higher than that using the squared return proxy. Also, after adjusting the variance forecasts for model risk, the degree of predictability of volatility models improves as evidenced by an increase in the values of adjusted R^2 of the MZ regressions.

In addition, applying our proposed methodology to several asset classes, we identify the models which are most affected by model risk, and find that volatility models carry a higher level of model risk during stressed market states, as expected. We also show that model misspecification risk generally contributes more to the total model risk than parameter estimation risk. It would be of interest to consider the model risk of volatility models in a multivariate setting as well as the model risk of market risk models.

Appendix A. Density functions for error distributions

Normal density function: For mean μ_z and standard deviation σ_z of z, this is given as:

$$f(z|\mu_z, \sigma_z) = \frac{1}{\sigma_z\sqrt{2\pi}} e^{-(z-\mu_z)^2/2\sigma_z^2}.$$

Student's *t* density function: It is written as, where ν denotes the degrees of freedom and $\Gamma(\cdot)$ denotes the Gamma function:

$$f(z|\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\nu\pi}} \frac{1}{(1+\frac{z^2}{\nu})^{\frac{\nu+1}{2}}}.$$

Skewed Student's *t* density function: Following Hansen (1994), it is given as:

$$f(z|\nu,\lambda) = \begin{cases} bc\left(1+\frac{1}{\nu-2}\left(\frac{bz+a}{1-\lambda}\right)^2\right)^{-(\nu+1)/2}, & \text{if } z < -a/b, \\ bc\left(1+\frac{1}{\nu-2}\left(\frac{bz+a}{1+\lambda}\right)^2\right)^{-(\nu+1)/2}, & \text{if } z \geq -a/b, \end{cases}$$

where the degree of freedom parameter ν with $2 < \nu < \infty$ controls the kurtosis and the skewness parameter λ is $-1 < \lambda < 1$. The constants a, b and c are given by:

$$a = 4\lambda c\left(\frac{\nu-2}{\nu-1}\right), \quad b^2 = 1 + 3\lambda^2 - a^2, \quad \text{and } c = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi(\nu-2)}\Gamma(\frac{\nu}{2})}.$$

Generalized error distribution (GED) density function: The probability density function of the generalized error distribution of the standardized residuals z beyond the threshold u is shown as below, where ξ and β are the shape and scale parameters with $\beta > 0$, respectively:

$$f(z|\xi,\beta) = \begin{cases} 1-(1+\xi z/\beta)^{-1/\xi}, & \text{if } \xi > 0, \\ 1-e^{-z/\beta}, & \text{if } \xi = 0, \end{cases} \quad for\ all\ z \geq u.$$

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We are grateful to editors, two anonymous referees and Michael P. Clements for their helpful feedback and comments. We also thank the participants of the 14th International Conference on Computational and Financial Econometrics (CFE 2020, Virtual). We acknowledge that this research received no external funding.

References

Alizadeh, S., Brandt, M. W., and Diebold, F. X. (2002). Range-based estimation of stochastic volatility models. *Journal of Finance*, 57(3):1047–1091.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625.

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

Aven, T. (2016). Risk assessment and risk management: Review of recent advances on their foundation. *European Journal of Operational Research*, 253(1):1–13.

Barndorff-Nielsen, O. E. and Shephard, N. (2002). Estimating quadratic variation using realized variance. *Journal of Applied Econometrics*, 17(5):457–477.

Barrieu, P. and Scandolo, G. (2015). Assessing financial model risk. *European Journal of Operational Research*, 242(2):546–556.

Basel Committee on Banking Supervision (2019). Minimum capital requirements for market risk. Retrieved from: `https://www.bis.org/bcbs/publ/d457.pdf(accessed08June2021)`.

Bauwens, L., Hafner, C. M., and Laurent, S. (2012). *Handbook of volatility models and their applications*, volume 3. John Wiley & Sons.

Blasques, F., Lucas, A., and van Vlodrop, A. C. (2021). Finite Sample Optimality of Score-Driven Volatility Models: Some Monte Carlo Evidence. *Econometrics and Statistics*, 19:47–57.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Bollerslev, T., Patton, A. J., and Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, 192(1):1–18.

Bollerslev, T. and Wright, J. H. (2001). High-frequency data, frequency domain inference, and volatility forecasting. *Review of Economics and Statistics*, 83(4):596–602.

Boucher, C. M., Daníelsson, J., Kouontchou, P. S., and Maillet, B. B. (2014). Risk models-at-risk. *Journal of Banking & Finance*, 44:72–92.

Danielsson, J., James, K. R., Valenzuela, M., and Zer, I. (2016). Model risk of risk models. *Journal of Financial Stability*, 23:79–91.

Daníelsson, J., Jorgensen, B. N., Samorodnitsky, G., Sarma, M., and de Vries, C. G. (2013). Fat tails, VaR and subadditivity. *Journal of Econometrics*, 172(2):283–291.

Dette, H., Golosnoy, V., and Kellermann, J. (2022). Correcting Intraday Periodicity Bias in Realized Volatility Measures. *Econometrics and Statistics*, 23:36–52.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3):253–263.

Emond, E. J. and Mason, D. W. (2002). A new rank correlation coefficient with application to the consensus ranking problem. *Journal of Multi-Criteria Decision Analysis*, 11(1):17–28.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, 50(4):987–1007.

Engle, R. F. and Siriwardane, E. N. (2018). Structural GARCH: the volatility-leverage connection. *Review of Financial Studies*, 31(2):449–492.

European Banking Authority (2014). Guidelines on common procedures and methodologies for the supervisory review and evaluation process (srep). Available online at: `https://eba.europa.eu/regulation-and-policy/supervisory-review-and-evaluation-srep-and-pillar-2/` (accessed 22 April 2021).

Federal Reserve (2011). Supervisory guidance on model risk management. Available online at: `https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm` (accessed 22 April 2021).

Forsberg, L. and Ghysels, E. (2007). Why do absolute returns predict volatility so well? *Journal of Financial Econometrics*, 5(1):31–67.

Fresoli, D. E. and Ruiz, E. (2016). The uncertainty of conditional returns, volatilities and correlations in dcc models. *Computational Statistics & Data Analysis*, 100:170–185.

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

Glasserman, P. and Xu, X. (2014). Robust risk measurement and model risk. *Quantitative Finance*, 14(1):29–58.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48(5):1779–1801.

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

Gourieroux, C. and Monfort, A. (2021). Model risk management: Valuation and governance of pseudo-models. *Econometrics and Statistics*, 17:1–22.

Green, T. C. and Figlewski, S. (1999). Market risk and model risk for a financial institution writing options. *Journal of Finance*, 54(4):1465–1499.

Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review*, 35:705–730.

Hansen, P. R. and Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1, 1)? *Journal of Applied Econometrics*, 20(7):873–889.

Hansen, P. R. and Lunde, A. (2006). Consistent ranking of volatility models. *Journal of Econometrics*, 131(1):97–121.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Huang, X. and Tauchen, G. (2005). The relative contribution of jumps to total price variance. *Journal of Financial Econometrics*, 3(4):456–499.

Jorion, P. (1995). Predicting volatility in the foreign exchange market. *Journal of Finance*, 50(2):507–528.

Kerkhof, J., Melenberg, B., and Schumacher, H. (2010). Model risk and capital reserves. *Journal of Banking & Finance*, 34(1):267–279.

Li, J. and Xiu, D. (2016). Generalized method of integrated moments for high-frequency data. *Econometrica*, 84(4):1613–1633.

McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative risk management: Concepts, techniques and tools - Revised edition.* Princeton university press.

Mincer, J. A. and Zarnowitz, V. (1969). The evaluation of economic forecasts. In *Economic forecasts and expectations: Analysis of forecasting behavior and performance*, pages 3–46. NBER.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica: Journal of the Econometric Society*, 59(2):347–370.

Pascual, L., Romo, J., and Ruiz, E. (2006). Bootstrap prediction for returns and volatilities in GARCH models. *Computational Statistics & Data Analysis*, 50(9):2293–2312.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256.

Takahashi, M., Watanabe, T., and Omori, Y. (2021). Forecasting Daily Volatility of Stock Price Index Using Daily Returns and Realized Volatility. *Econometrics and Statistics*.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64:1067–1084.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 48(4):817–838.