Two are better than one: volatility forecasting using multiplicative component GARCH models*

Christian Conrad† and Onno Kleen‡

February 28, 2018

PRELIMINARY DRAFT

Abstract

We examine the forecast performance of multiplicative volatility models that can be decomposed into a short- and a long-term component. First, we show that in multiplicative models, returns have a higher kurtosis and squared returns have a more persistent autocorrelation function than in the nested GARCH model. Second, we provide theoretical and simulation evidence suggesting that the QLIKE loss should be preferred relative to the squared errors loss when comparing volatility forecasts. In a Monte-Carlo simulation, we investigate how the multiplicative structure affects forecast performance both in comparison to the nested GARCH model and the popular HAR model. Finally, we consider an application to S&P 500 returns. Based on the QLIKE loss and forecast horizons of two- to three-months ahead, our results show that multiplicative GARCH models incorporating financial and macroeconomic variables improve upon the HAR model.

Keywords: Forecast evaluation, GARCH-MIDAS, Mincer-Zarnowitz regression, volatility persistence, volatility component model, long-term volatility, model confidence set.

JEL Classification: C53, C58, G12

*We would like to thank Richard Baillie, Matei Demetrescu, Markus Haas, Karin Stürmer, Timo Teräsvirta, and Peter Winker for helpful comments. This research was supported by a travel grant of the International Association of Applied Econometrics.

†Christian Conrad, Department of Economics, Heidelberg University, Bergheimer Strasse 58, 69115 Heidelberg, Germany, Email: christian.conrad@awi.uni-heidelberg.de; Phone: +49 6221 54 3173.

‡Onno Kleen, Department of Economics, Heidelberg University, Bergheimer Strasse 58, 69115 Heidelberg, Germany, Email: onno.kleen@awi.uni-heidelberg.de; Phone: +49 6221 54 2930.

1 Introduction

The idea to model volatility as consisting of multiple components has a long tradition in financial econometrics (see, for example, Ding and Granger, 1996, and Engle and Lee, 1999). However, these early models typically featured additive volatility components and did not allow for explanatory variables in the conditional variance. More recently, the focus of attention has shifted to multiplicative component models (see, for example, Engle and Rangel, 2008, Engle et al., 2013, Amado and Teräsvirta, 2013, 2017, and Han and Kristensen, 2015). Specifically, the class of GARCH-MIDAS models proposed in Engle et al. (2013) has been proven to be extremely useful for analyzing the link between financial conditions and the macroeconomic environment (see Asgharian et al., 2013, Conrad and Loch, 2015, and Dorion, 2016). In the GARCH-MIDAS, a unit-variance GARCH component fluctuates around a smoothly time-varying long-term component that is a function of (macroeconomic) explanatory variables. By allowing for a mixed-frequency setting, this approach bridges the gap between daily stock returns and low-frequency (e.g., monthly, quarterly) explanatory variables. For further applications of GARCH-MIDAS-type models to financial time series such as stock price indices, individual stock prices or oil prices see, for example, Conrad et al. (2014), Opschoor et al. (2014), Dominicy and Vander Elst (2015), Lindblad (2017), Amendola et al. (2017) and Pan et al. (2017). For a recent survey on multiplicative component models see Amado et al. (2018).

Our contribution to this recent strand of literature is twofold. First, we analyze some statistical properties of multiplicative component GARCH models (M-GARCH) that did not receive attention so far. Assuming a two-component structure with one component being a GJR-GARCH (see Glosten et al., 1993), we show that the kurtosis of the two-component model is always bigger than the kurtosis of the nested GJR-GARCH component. The autocorrelation function (ACF) of the squared returns is shown to be a weighted average of the ACF of the long-term component and the ACF of the GJR-GARCH component. If the long-term component is sufficiently persistent, it dominates the behavior of the ACF of the overall process and can mimic long-memory type dynamics. Both findings suggest a multiplicative component structure in the volatility of stock returns as a potential explanation for the common failure of simple one-component GARCH models in adequately capturing the leptokurticity and extreme volatility persistence in observed returns. It should also be noted that our results are remarkably similar to the recent findings in Han (2015) on GARCH-X models, even though Han (2015) considers models with an additive explanatory variable in the conditional variance and focuses on the asymptotic limit of the sample kurtosis and the sample ACF. Finally, it is important to highlight that although our results on the kurtosis and the ACF are presented for a GJR-GARCH(1,1) short-term component, they directly extend to a covariance

stationary GJR-GARCH(p, q) component.

Further, we analyze the usefulness of the squared error (SE) loss and the QLIKE loss for evaluating volatility forecasts. Our insights suggest that the QLIKE is the more appropriate measure because – in contrast to the SE loss – the QLIKE loss does not depend on the prevalent volatility regime. We thus complement the arguments put forth in Patton (2011) and Brownlees et al. (2012) in favor of the QLIKE. Based on the analysis of the SE loss, we derive an upper bound for the population R^2 in the k-step ahead Mincer-Zarnowitz regression of the squared return on the volatility forecast. Our result nests the case of a simple GARCH(1,1) that was discussed in Andersen and Bollerslev (1998). We derive an explicit expression for the one-step ahead R^2 of the GARCH-MIDAS specification and show that the R^2 is increasing in the variability of the long-term component. This feature leads to the unpleasant property that the R^2 is particularly high in situations in which the SE loss is also high. Clearly, this finding questions the Mincer-Zarnowitz R^2 as a reasonable measure of forecast accuracy. In a Monte-Carlo simulation, we provide further evidence for our theoretical findings. In addition, the simulation shows that a two-component structure in the data generating process (DGP) leads to the well known IGARCH effect when a simple one-component model is applied to the data and the long-term component is of sufficient importance. Finally, we compare the forecast accuracy of the two-component model with the one of the popular Heterogeneous Autoregressive (HAR) model of Corsi (2009) when again the true DGP has a two-component structure. The HAR model is based on additive components of the realized volatility of the previous day, the previous week and the previous month and, thus, has a similar motivation as multiplicative models with short- and long-term components. Our results show that the correctly specified M-GARCH models clearly outperform the HAR model in terms of out-of-sample forecast performance. Hence, a direct modeling of the realized variance does not appear to be a good substitute for a proper identification of the true component structure. The simulations also show that the power of Diebold-Mariano tests for equal forecast accuracy is higher under the QLIKE loss than under the SE loss. This effect is particularly strong in high-volatility regimes. Again this findings supports the arguments in Patton and Sheppard (2009) and Patton (2011) in favor of the QLIKE.

Second, we apply the GARCH-MIDAS model to a long time series of S&P 500 returns combined with data on U.S. macroeconomic and financial conditions. The empirical application focuses on certain aspects that have remained unexplored in previous applications of GARCH-MIDAS models. While previous studies have either used ex-post or first release data for the full (in-)sample period (see, for example, Engle et al., 2013, and Conrad and Loch, 2015), we base the estimates of our GARCH-MIDAS model on a rolling window of vintage data that was available in real-time. To the best of our knowledge, Lindblad (2017) appears to be the only other paper that makes use

of real-time data when estimating GARCH-MIDAS models. Using real-time data is of particular importance when employing macroeconomic time series which are often revised substantially after the first release. For example, we find that for industrial production the average absolute deviation of the estimate of the long-term component based on final data for the full-sample from the rolling-window estimate based on real-time data is about 8%. By using real-time data, our study does not suffer from a 'look-ahead-bias' and allows for a realistic evaluation of the true out-of-sample forecasting ability of the different models. Instead of comparing the GARCH-MIDAS forecasts with the forecasts from the nested GARCH model (see Asgharian et al., 2013, and Lindblad, 2017) or a GARCH-MIDAS with realized volatility as the explanatory variable (see Conrad and Loch, 2015), we consider the empirically 'hard-to-beat' HAR model as the main competitor. Further, instead of comparing each GARCH-MIDAS model separately with a benchmark model, we evaluate all models at once by constructing model confidence sets (MCS) as introduced in Hansen et al. (2011). Based on the QLIKE as our favorite measure of forecast accuracy, we find that the MCS for the one-day to one-month ahead forecasts consists primarily of a GARCH-MIDAS model based on the VIX as the explanatory variable and the HAR model. For forecast horizons of two- and three-months ahead, GARCH-MIDAS models based on the National Financial Conditions Index (NFCI) and housing starts are the preferred specifications. Thus, our empirical results suggest that the best GARCH-MIDAS specifications are at least as good as the HAR model in short-term forecasting and even superior in long-term forecasting.

To facilitate the replication of our results, we provide R packages for downloading real-time data from the ALFRED database of the Federal Reserve Bank of St. Louis (see Kleen, 2017) as well as for estimating (mixed-frequency) M-GARCH models (see Kleen, 2018).[1]

Our paper is organized as follows. In Section 2, the model and our theoretical results are presented. In Section 3, we perform a simulation study and, in Section 4, we apply the M-GARCH model to S&P 500 return data. Section 5 concludes. All proofs are deferred to Appendix A.

2 The Multiplicative Component GARCH Model

In this section, the M-GARCH model is introduced and its theoretical properties are derived. In particular, we show that the M-GARCH model inherits certain time series properties that are in line with stylized facts typically observed for financial return data but that cannot be captured by simple GARCH models.

[1]The packages are available at: *cran.r-project.org/package=alfred* and *cran.r-project.org/package=mfGARCH*.

2.1 Model specification

We denote daily returns by $\varepsilon_{i,t}$, whereby t refers to a certain period (e.g. a week or a month) and $i = 1, \ldots, I_t$ to days within that period. Daily (demeaned) returns are given by

$$\varepsilon_{i,t} = \sigma_{i,t} Z_{i,t}, \tag{1}$$

where $Z_{i,t}$ is an *i.i.d.* innovation process with mean zero and variance one. Let $\mathcal{F}_{i,t}$ denote the information set up to day i in period t and define $\mathcal{F}_t := \mathcal{F}_{I_t,t}$. As we discuss below, we will assume that $\sigma^2_{i,t}$ is measurable with respect to $\mathcal{F}_{i-1,t}$. Hence, $\sigma^2_{i,t}$ represents the conditional variance of returns, i.e. $\mathbf{Var}(\varepsilon_{i,t}|\mathcal{F}_{i-1,t}) = \sigma^2_{i,t}$. In the M-GARCH model, we specify the conditional variance as the product of a short-term and a long-term component. That is, we decompose $\sigma^2_{i,t}$ as follows:

$$\sigma^2_{i,t} = g_{i,t}\tau_t. \tag{2}$$

We refer to $g_{i,t}$ as the short-term component and to τ_t as the long-term component. While $g_{i,t}$ changes at the daily frequency, τ_t is constant across all days within period t and, instead, changes at the low frequency only.

The short-term component is intended to describe the well known day-to-day clustering of volatility and is assumed to follow a mean-reverting unit-variance GJR-GARCH(1,1) process:

$$\begin{aligned} g_{i,t} &= (1 - \alpha - \gamma/2 - \beta) + \left(\alpha + \gamma \mathbb{1}_{\{\varepsilon_{i-1,t}<0\}}\right) \frac{\varepsilon^2_{i-1,t}}{\tau_t} + \beta g_{i-1,t} \\ &= \sum_{j=0}^{\infty} \beta^j \left((1 - \alpha - \gamma/2 - \beta) + (\alpha + \gamma \mathbb{1}_{\{\varepsilon_{i-j,t}<0\}}) Z^2_{i-1,t}\right), \end{aligned} \tag{3}$$

where the second line makes clear that $g_{i,t}$ can be entirely written in terms of past values of $Z^2_{i,t}$. We make the following assumptions regarding the innovation $Z_{i,t}$ and the parameters of the short-term component.

Assumption 1. *Let $Z_{i,t}$ be i.i.d. with $\mathbf{E}[Z_{i,t}] = 0$ and $\mathbf{E}[Z^2_{i,t}] = 1$. Further, $Z^2_{i,t}$ is assumed to have a nondegenerate distribution and $\kappa = \mathbf{E}[Z^4_{i,t}] < \infty$.*

Assumption 2. *We assume that $\alpha > 0$, $\alpha + \gamma > 0$, $\beta \geq 0$ and $\alpha + \gamma/2 + \beta < 1$. Moreover, we assume that $(\alpha + \gamma/2)^2\kappa + 2(\alpha + \gamma/2)\beta + \beta^2 < 1$.*

Note that Assumption 1 implies that $\kappa > 1$. Assumptions 1 and 2 imply that $\varepsilon_{i,t}/\sqrt{\tau_t} = \sqrt{g_{i,t}} Z_{i,t}$ is a covariance stationary GJR-GARCH(1,1) process. The first- and second-order moment of $g_{i,t}$

are given by $\mathbf{E}[g_{i,t}] = 1$,

$$\mathbf{E}[g_{i,t}^2] = \frac{1 - (\alpha + \gamma/2 + \beta)^2}{1 - (\alpha + \gamma/2)^2\kappa - 2(\alpha + \gamma/2)\beta - \beta^2} \tag{4}$$

and the fourth moment of $\sqrt{g_{i,t}}Z_{i,t}$ is finite.

The second component, τ_t, should be thought of as describing smooth movements in the conditional variance. Instead of explicitly specifying τ_t, we will simply assume that it follows a covariance stationary process.

Assumption 3. *The long-term component $\tau_t > 0$ is covariance stationary and measurable with respect to $\mathcal{F}_{t-1}$. Moreover, we assume that τ_t and $Z_{i,t-j}$ are independent for all t, i and j.*

Assuming independence of τ_t and $Z_{i,t+j}$ is stronger than what is needed for both estimating M-GARCH models (see Wang and Ghysels, 2015) as well as testing for an omitted long-term component in one-component GARCH models (see Conrad and Schienle, 2018). We assume independence because it will allow us to obtain simple expressions for the kurtosis of the returns, the autocorrelation function of the squared returns and for the comparison of alternative measures of forecast accuracy.[2] Alternatively, we could, for example, assume that τ_t is strictly exogenous with respect to the first four moments of $Z_{i,t}$.

Assumptions 1, 2 and 3 imply that the $\varepsilon_{i,t}$ have mean zero, are uncorrelated and have an unconditional variance given by $\mathbf{Var}(\varepsilon_{i,t}) = \mathbf{E}[\tau_t]$. Moreover, the unconditional variance of the squared returns is well-defined, $\mathbf{Var}(\varepsilon_{i,t}^2) = \kappa\mathbf{E}[\tau_t^2]\mathbf{E}[g_{i,t}^2] - \mathbf{E}[\tau_t]^2$.

2.2 Related specifications

The model described in Section 2.1 nests the GARCH-MIDAS of Engle et al. (2013). In the GARCH-MIDAS the long-term component is specified as a function of lagged values of an exogenous explanatory variable X_t. That is, τ_t may be written as $\tau_t = f(X_{t-1}, X_{t-2}, \ldots, X_{t-K})$ with $f(\cdot) > 0$.[3] The most common specification of the long-term component expresses the natural logarithm of τ_t as a weighted sum of the $K \geq 1$ lagged values of X_t,

$$\log \tau_t = m + \theta \sum_{l=1}^{K} \varphi_l(w_1, w_2) X_{t-l}, \tag{5}$$

[2]Essentially, Han and Kristensen (2015) make the same independence assumption in their multiplicative GARCH model.

[3]While we focus on multiplicative GARCH models, Han and Park (2014) and Han (2015) analyze the properties of a GARCH-X specification with an explanatory variable that enters additively into the conditional variance equation. See also Han and Kristensen (2014) and Francq and Thieu (2015).

where the weights have a Beta polynomial structure,

$$\varphi_l(w_1, w_2) = \frac{(l/(K+1))^{w_1-1} \cdot (1 - l/(K+1))^{w_2-1}}{\sum_{j=1}^{K} (j/(K+1))^{w_1-1} \cdot (1 - j/(K+1))^{w_2-1}}. \tag{6}$$

By construction, the weights sum up to one, i.e. $\sum_{l=1}^{K} \varphi_l(w_1, w_2) = 1$.

A measure that is often used to quantify the relative importance of the long-term component is the following variance ratio:

$$VR = \frac{\mathbf{Var}(\log(\tau_t))}{\mathbf{Var}(\log(\tau_t g_t))}, \tag{7}$$

where $g_t = \sum_{i=1}^{I_t} g_{i,t}$. The ratio measures how much of the total variation in the (log) conditional variance can be explained by the variation in the (log) long-term component.

Although the model described in Section 2.1 involves mixed-frequency data, we can rewrite the returns as $\varepsilon_s = \varepsilon_{i,t}$, where the new daily index s is defined as $s = s(i,t) = \sum_{l=0}^{t-1} I_l + i$ with $I_0 = 0$. If the length of period t is simply one day, both components vary at the same frequency and our model coincides with the specification discussed in Conrad and Schienle (2018). The latter model nests the component specification considered in Wang and Ghysels (2015) when $f(\cdot)$ is linear in the lagged realized variances.

Finally, note that by multiplying eq. (3) by τ_t, we obtain

$$\sigma_{i,t}^2 = \tau_t(1 - \alpha - \gamma/2 - \beta) + \left(\alpha + \gamma \mathbb{1}_{\{\varepsilon_{i-1,t}<0\}}\right) \varepsilon_{i-1,t}^2 + \beta \sigma_{i-1,t}^2. \tag{8}$$

Thus, eq. (1) and (8) can be viewed an asymmetric GARCH process with time-varying intercept. In a similar manner, Baillie and Morana (2009) introduced the adaptive FIGARCH model which assumes that $\sigma_{i,t}^2$ follows a FIGARCH with time-varying intercept. However, in their model the time-varying intercept purely depends on time. If $\tau_t = \omega/(1 - \alpha - \gamma/2 - \beta)$ is constant, our model reduces to the GJR-GARCH with intercept ω.

2.3 Kurtosis

Financial returns are often found to be leptokurtic. As usual, we measure leptokurticity by means of the kurtosis coefficient. Under Assumptions 1, 2 and 3, the kurtosis of the returns defined in eq. (1) is given by

$$\mathcal{K}^{MG} = \frac{\mathbf{E}[\varepsilon_{i,t}^4]}{(\mathbf{E}[\varepsilon_{i,t}^2])^2} = \frac{\mathbf{E}[\sigma_{i,t}^4]}{(\mathbf{E}[\sigma_{i,t}^2])^2} \kappa > \kappa.$$

Thus, the kurtosis of the M-GARCH process is larger than the kurtosis of the innovation $Z_{i,t}$. This is a well known feature of GARCH-type processes. The following proposition relates the kurtosis

$\mathcal{K}^{MG}$ of the M-GARCH to the kurtosis $\mathcal{K}^{GA}$ of the nested GARCH(1,1).

Proposition 1. *Under Assumptions 1, 2 and 3, the kurtosis $\mathcal{K}^{MG}$ of an M-GARCH process is given by*

$$\mathcal{K}^{MG} = \frac{\mathbf{E}[\tau_t^2]}{\mathbf{E}[\tau_t]^2} \cdot \mathcal{K}^{GA} \geq \mathcal{K}^{GA},$$

where $\mathcal{K}^{GA} = \kappa \cdot \mathbf{E}[g_{i,t}^2]$ is the kurtosis of the nested GARCH process and where the equality holds if and only if τ_t is constant.

Hence, for non-constant τ_t the kurtosis $\mathcal{K}^{MG}$ is the product of $\mathcal{K}^{GA}$ and the ratio $\mathbf{E}[\tau_t^2]/\mathbf{E}[\tau_t]^2 > 1$. When $\tau_t = \omega/(1-\alpha-\gamma/2-\beta)$ is constant, Proposition 1 nests the kurtosis of the asymmetric GARCH model. Thus, for highly volatile long-term components the kurtosis of an M-GARCH process is much larger than the kurtosis of the nested GARCH model.[4] Moreover, when estimating a GARCH model, it is often assumed that $Z_{i,t}$ is standard normal so that $\kappa = 3$. Our result may explain why in empirical applications the 'deGARCHed' residuals, $\varepsilon_{i,t}/\sqrt{g_{i,t}}$, often still exhibit excess kurtosis. In the multiplicative model the kurtosis of $\varepsilon_{i,t}/\sqrt{g_{i,t}}$ is given by $3 \cdot \mathbf{E}[\tau_t^2]/\mathbf{E}[\tau_t]^2 > 3$.

2.4 Autocorrelation function

Empirically, the ACF of squared returns is often found to decay more slowly than the exponentially decaying ACF implied by the simple GARCH(1,1) model. In the literature on GARCH models, this is usually interpreted as either evidence for long-memory (see, e.g., Baillie et al., 1996), structural breaks (see, e.g., Hillebrand, 2005) or an omitted persistent covariate (see Han and Park, 2014) in the conditional variance.

The following proposition shows that the theoretical ACF of the M-GARCH process has a much slower decay than the ACF of the nested GARCH component if the long-term component is sufficiently persistent. Hence, the multiplicative structure provides an alternative explanation for the empirical observation of highly persistent ACFs of squared returns. For simplicity in the notation, we consider the case that both components are varying at the same frequency, i.e the length of the period t is one day.

Proposition 2. *If the long-term component is observed at the daily frequency and Assumptions 1, 2 and 3 are satisfied, the ACF, ρ_k^{MG}, of an M-GARCH process is given by*

$$\rho_k^{MG} = \mathbf{Corr}(\varepsilon_s^2, \varepsilon_{s-k}^2) \quad = \quad \rho_k^{\tau} \frac{\mathbf{Var}(\tau_s)}{\mathbf{Var}(r_s^2)} + \rho_k^{GA} \frac{\left(\rho_k^{\tau}\mathbf{Var}(\tau_s) + \mathbf{E}[\tau_s]^2\right)\mathbf{Var}(g_s Z_s^2)}{\mathbf{Var}(r_s^2)} \tag{9}$$

[4] Han (2015) obtains a similar result for the sample kurtosis of the returns from a GARCH-X model with a covariate that can be either stationary or non-stationary.

with $\rho_k^\tau = \mathbf{Corr}(\tau_s, \tau_{s-k})$ *and*

$$\rho_k^{GA} = \mathbf{Corr}(g_s Z_s^2, g_{s-k} Z_{s-k}^2) = (\alpha + \gamma/2 + \beta)^{k-1} \frac{(\alpha + \gamma/2)(1 - (\alpha + \gamma/2)\beta - \beta^2)}{1 - 2(\alpha + \gamma/2)\beta - \beta^2}$$

being the ACF of the GJR-GARCH component.[5]

Proposition 2 shows that the ACF of the M-GARCH model is given by the sum of two terms: The first term is given by the ACF of the long-term component ρ_k^τ times a constant, whereas the second term equals the exponentially decaying ACF of the nested GARCH model ρ_k^{GA} times a ratio that depends again on ρ_k^τ. Hence, if τ_s is sufficiently persistent, ρ_k^{MG} will essentially behave as ρ_k^τ for k large.[6] For τ_s being constant, the first term in eq. (9) is equal to zero and the second term reduces to the ACF of an asymmetric GARCH(1,1). Also, note that the ratio $\mathbf{Var}(\tau_s)/\mathbf{Var}(r_s^2)$ is closely related to the variance ratio defined in eq. (7) and essentially measures the importance of the long-term component.

The implications of Proposition 2 are depicted in Figure 1. Due to the additive structure of the correlation function, a highly persistent exogenous covariate causes the ACF of the M-GARCH model (solid line) to have a decay pattern that is quite different from the ACF of the nested GARCH component (dashed line). The ACF of the M-GARCH behaves like the ACF typically observed for squared returns, i.e. it decreases exponentially at first but remains larger than zero for high lags. Figure 1 shows that the second term (i.e. the ACF of the GARCH component) in eq. (9) determines the decay behavior of ρ_k^{MG} when k is small, while the first term dominates when k is large.

[Figure 1 here]

As for the kurtosis, our result may explain why in empirical applications the squared de-GARCHed residuals, ε_s^2/g_s, are still substantially autocorrelated. In the multiplicative model, the ACF of these residuals is given by $\rho_k^\tau \cdot \mathbf{Var}(\tau_s)/(\kappa \mathbf{E}[\tau_s^2] - \mathbf{E}[\tau_s]^2)$, which follows the rate of decay of the long-term component.

2.5 Forecast evaluation and Mincer-Zarnowitz regression

In this section, we discuss the properties of the expected squared error (SE) loss, the QLIKE loss as well as the Mincer-Zarnowitz R^2 as measures of forecast accuracy when forecasts from M-GARCH models are evaluated against squared returns as a proxy for the true but latent volatility.

[5] Note that ρ_k^{GA} reduces to the ACF of a (symmetric) GARCH(1,1) when $\gamma = 0$ (see Karanasos, 1999).

[6] Again, Han (2015) also obtains a two component structure for the sample ACF of the squared returns from a GARCH-X model with a fractionally integrated covariate.

We assume that forecasts are produced at the last day I_t of period t. For the time being, we consider the k-step ahead volatility forecast coming from *some* volatility model (not necessarily from the component GARCH model) and denote it by $h_{k,t+1|t}$ with $k \leq I_{t+1}$. If the model under consideration is the multiplicative component model then $h_{k,t+1|t} = \tau_{t+1} g_{k,t+1|t}$, where $g_{k,t+1|t} = \mathbf{E}[g_{k,t+1}|\mathcal{F}_t] = 1 + (\alpha + \gamma/2 + \beta)^{k-1}(g_{1,t+1|t} - 1)$. When evaluating the volatility forecast, one has to deal with the problem that the true conditional variance, $\sigma^2_{k,t+1}$, is unobservable. Patton (2011) discusses the situation in which the forecast evaluation is based on some conditionally unbiased volatility proxy $\hat{\sigma}^2_{k,t+1}$ instead. He defines a loss function $L(\sigma^2_{k,t+1}, h_{k,t+1|t})$ as *robust* if the expected loss ranking of two competing forecasts is preserved when replacing $\sigma^2_{k,t+1}$ by $\hat{\sigma}^2_{k,t+1}$. As shown in Patton (2011), the SE loss function,

$$\text{SE}(\sigma^2_{k,t+1}, h_{k,t+1|t}) = (\sigma^2_{k,t+1} - h_{k,t+1|t})^2 \tag{10}$$

as well as the QLIKE loss,

$$\text{QLIKE}\left(\sigma^2_{k,t+1}, h_{k,t+1|t}\right) = \frac{\sigma^2_{k,t+1}}{h_{k,t+1|t}} - \ln\left(\frac{\sigma^2_{k,t+1}}{h_{k,t+1|t}}\right) - 1, \tag{11}$$

belong to the class of robust loss functions. A necessary condition for a loss function to be robust is that the optimal forecast is the conditional mean of the variable to be forecasted. This rules out, for example, the mean absolute error loss. Further, the SE is the only robust loss function that depends solely on the forecast error, $\sigma^2_{k,t+1} - h_{k,t+1|t}$, while the QLIKE is the only robust loss function that depends solely on the standardized forecast error, $\sigma^2_{k,t+1}/h_{k,t+1|t}$ (see Patton, 2011). Comparing the conditional variance of the forecast error and the standardized forecast error suggests that the SE is much more sensitive with respect to extreme observations in the sample. Similarly, it can be shown that the moment conditions required for Diebold and Mariano (1995) tests are much stronger under SE loss than under QLIKE loss (see Patton, 2006).

Finally, note that the ranking obtained from the SE and the QLIKE for two competing models should be the same as long as the two models are "correctly specified, free from estimation error, and when the information sets of one of the forecasters nests the other" (Patton, 2016). However, if one of these conditions is violated, the SE and QLIKE can lead to differing model rankings. In this case, it is up to the researcher to decide which loss function to prefer. As the subsequent discussion of the SE loss and the QLIKE loss shows, the QLIKE loss might be the more appropriate choice in our setting.

2.5.1 Squared error loss

Consider the model given by eq. (1) and the corresponding expected SE from evaluating a variance forecast $h_{k,t+1|t}$ against the noisy but conditionally unbiased proxy $\hat{\sigma}^2_{k,t+1} = \varepsilon^2_{k,t+1}$.[7] Using that $\mathbf{E}[\varepsilon^2_{k,t+1}|\mathcal{F}_{k-1,t+1}] = \sigma^2_{k,t+1}$, it is straightforward to show that

$$\mathbf{E}[\mathrm{SE}(\varepsilon^2_{k,t+1}, h_{k,t+1|t})] = \mathbf{E}[\mathrm{SE}(\sigma^2_{k,t+1}, h_{k,t+1|t})] + (\kappa - 1)\mathbf{E}[\sigma^4_{k,t+1}]. \tag{12}$$

That is, the expected SE based on the noisy proxy equals the expected SE based on the latent volatility plus a term that depends on the fourth moment, κ, of $Z_{i,t}$ and the expected value of the squared conditional variance. Hence, using a noisy proxy for forecast evaluation can lead to a substantially higher expected SE than based on the latent volatility. Patton (2011, p.248) basically made the same point by arguing that "although the ranking obtained from a robust loss function will be invariant to noise in the proxy, the actual *level* of expected loss obtained using a proxy will be larger than that which would be obtained when using the true conditional variance". For further illustration, consider the case of a one-day ahead forecast. In this case, $h_{1,t+1|t} = \tau_{t+1} g_{1,t+1|t}$ and thus $\mathbf{E}[\mathrm{SE}(\sigma^2_{1,t+1}, h_{1,t+1|t})] = 0$. Nevertheless, eq. (12) implies that the expected SE is still non-zero:

$$\mathbf{E}[\mathrm{SE}(\varepsilon^2_{1,t+1}, h_{1,t+1|t})] = (\kappa - 1)\mathbf{E}[\sigma^4_{1,t+1}] = (\kappa - 1)\mathbf{E}[\tau^2_{t+1}]\mathbf{E}[g^2_{1,t+1}]. \tag{13}$$

That is, although the forecaster is using the correct model, the expected SE can be large when $Z_{i,t}$ is leptokurtic or $\mathbf{E}[\sigma^4_{i,t}]$ is large. Brownlees et al. (2012) have made a similar point by arguing that the bias of $\mathbf{E}[\mathrm{SE}(\varepsilon^2_{1,t+1}, h_{1,t+1|t})]$ is proportional to the square of the true variance which makes it difficult to compare losses across different volatility regimes.

Next, we consider the expected SE conditional on the information available at time t. If the conditional variance forecast is *correctly specified*, i.e. $h_{k,t+1|t} = \tau_{t+1} g_{k,t+1|t}$, a conditional version of the first term in eq. (12) can be written as

$$\begin{aligned}
\mathbf{E}[\mathrm{SE}(\sigma^2_{k,t+1}, h_{k,t+1|t})|\mathcal{F}_t] &= \mathbf{E}[\tau^2_{t+1}(g_{k,t+1} - g_{k,t+1|t})^2|\mathcal{F}_t] \\
&= \tau^2_{t+1} \cdot \mathbf{E}\left[\left(\alpha \sum_{i=1}^{k-1} (\alpha + \gamma/2 + \beta)^{i-1} v_{k-i,t+1}\right)^2 |\mathcal{F}_t\right] \\
&= \tau^2_{t+1} \cdot (\kappa - 1)\alpha^2 \sum_{i=1}^{k-1} (\alpha + \gamma/2 + \beta)^{2(i-1)} \mathbf{E}[g^2_{k-i,t+1}|\mathcal{F}_t]
\end{aligned} \tag{14}$$

where $v_{k-i,t+1} = g_{k-i,t+1}(Z^2_{k-i,t+1} - 1)$. Equation (14) is given by the squared long-term component

[7] To illustrate the severeness of the noise, consider an example with $Z_{k,t+1} \sim \mathcal{N}(0,1)$. Then $\varepsilon^2_{k,t+1}$ will either over- or underestimate the true $\sigma^2_{k,t+1}$ by more than 50% with a probability of about 74%. In Section 3, we will consider the case that realized variance is used as a proxy for $\sigma^2_{k,t+1}$.

times the conditional mean squared error of the k-step ahead volatility prediction of a GARCH(1,1) (see, Baillie and Bollerslev, 1992). That is, although τ_{t+1} is a function of information available at t and, hence, does not need to be forecasted itself, it strongly affects the conditional SE loss. This observation illustrates that – even in the absence of noise on the volatility proxy – the level of the SE strongly depends on the volatility regime. Thus, even when we use precise volatility proxies such as realized volatility based on high-frequency intra-day data, we should expect that the outcome of tests for forecast accuracy is sensitive with respect to the volatility regime.

This property of the SE also applies when considering cumulative volatility forecasts. Again, we assume that forecasts are constructed at the last day of month t. The k-step ahead cumulative forecast is then given by

$$h_{1:k,t+1|t} \quad = \quad \tau_{t+1} \sum_{i=1}^{k} g_{i,t+1|t} = \tau_{t+1} \left(k + (g_{1,t+1} - 1) \frac{1 - (\alpha + \beta + \gamma/2)^k}{1 - \alpha - \beta - \gamma/2} \right). \tag{15}$$

As a proxy for the latent volatility during the first k days of period $t+1$, we use $\hat{\sigma}^2_{1:k,t+1} = \sum_{i=1}^{k} \varepsilon^2_{i,t+1}$. Again, it follows directly that

$$\mathbf{E}\left[\mathrm{SE}\left(\hat{\sigma}^2_{1:k,t+1}, h_{1:k,t+1|t} \right) \middle| \mathcal{F}_t \right] = \tau^2_{t+1} \cdot \mathbf{E}\left[\mathrm{SE}\left(\sum_{i=1}^{k} g_{i,t+1} Z^2_{i,t+1}, \sum_{i=1}^{k} g_{i,t+1|t} \right) \middle| \mathcal{F}_t \right] \tag{16}$$

depends on the squared long-term component.

2.5.2 QLIKE loss

Next, we turn to the QLIKE loss given in eq. (11). For simplicity, we directly consider the conditional expected QLIKE of the cumulative k-day ahead forecast which is given by

$$\begin{aligned}
\mathbf{E}\left[\mathrm{QLIKE}\left(\hat{\sigma}^2_{1:k,t+1}, h_{1:k,t+1|t} \right) \mathcal{F}_t \right] &= \mathbf{E}\left[\mathrm{QLIKE}\left(\sum_{i=1}^{k} \varepsilon^2_{i,t+1}, \tau_{t+1} \sum_{i=1}^{k} g_{i,t+1|t} \right) \middle| \mathcal{F}_t \right] \\
&= \mathbf{E}\left[\frac{\tau_{t+1} \sum_{i=1}^{k} g_{i,t+1}}{\tau_{t+1} \sum_{i=1}^{k} g_{i,t+1|t}} - \ln\left(\frac{\tau_{t+1} \sum_{i=1}^{k} g_{i,t+1} Z^2_{i,t+1}}{\tau_{t+1} \sum_{i=1}^{k} g_{i,t+1|t}} \right) - 1 \middle| \mathcal{F}_t \right] \\
&= \mathbf{E}\left[\frac{\sum_{i=1}^{k} g_{i,t+1} Z^2_{i,t+1}}{\sum_{i=1}^{k} g_{i,t+1|t}} - \ln\left(\frac{\sum_{i=1}^{k} g_{i,t+1} Z^2_{i,t+1}}{\sum_{i=1}^{k} g_{i,t+1|t}} \right) - 1 \middle| \mathcal{F}_t \right] \\
&= \mathbf{E}\left[\mathrm{QLIKE}\left(\sum_{i=1}^{k} g_{i,t+1} Z^2_{i,t+1}, \sum_{i=1}^{k} g_{i,t+1|t} \right) \middle| \mathcal{F}_t \right].
\end{aligned}$$

In contrast to the conditional expected SE, see eq. (16), we find that the conditional expected QLIKE loss does not depend on τ_{t+1}. Indeed, the conditional expected QLIKE of the M-GARCH volatility forecast reduces to the conditional expected QLIKE of the nested GARCH component.

Of course, when calculating the QLIKE based on estimated quantities, the actual QLIKE will depend on the ratio $\tau_{t+1}/\hat{\tau}_{t+1}$. Nevertheless, we should expect that the outcome of tests for forecast accuracy based on the QLIKE are less sensitive with respect to the volatility regime than tests based on the SE.

2.5.3 Mincer-Zarnowitz regression

Finally, we consider the problem of evaluating the k-step ahead volatility forecast by means of the coefficient of determination, R_k^2, of the following Mincer-Zarnowitz regression:

$$\varepsilon_{k,t+1}^2 = \delta_0 + \delta_1 h_{k,t+1|t} + \eta_{k,t+1}. \tag{17}$$

As shown in Hansen and Lunde (2006), the ranking of competing one-step-ahead volatility forecasts based on the R_1^2 of the Mincer-Zarnowitz regression is robust to using the proxy $\varepsilon_{1,t+1}^2$ instead of the latent conditional variance $\sigma_{1,t+1}^2$ as the dependent variable. We consider the situation in which $h_{k,t+1|t}$ is chosen optimally given available information, i.e. $h_{k,t+1|t} = \mathbf{E}[\sigma_{k,t+1}^2|\mathcal{F}_t]$. In this case, the population parameters of the Mincer-Zarnowitz regression are given by $\delta_0 = 0$ and $\delta_1 = 1$ and, hence, the population R_k^2 can be written as:

$$\begin{aligned} R_k^2 &= 1 - \frac{\mathbf{Var}(\eta_{k,t+1})}{\mathbf{Var}(\varepsilon_{k,t+1}^2)} = 1 - \frac{\mathbf{E}[\mathrm{SE}(\varepsilon_{k,t+1}^2, h_{k,t+1|t})]}{\mathbf{Var}(\varepsilon_{k,t}^2)} \\ &\leq 1 - \frac{(\kappa-1)\mathbf{E}[\sigma_{k,t}^4]}{\kappa\mathbf{E}[\sigma_{k,t}^4] - (\mathbf{E}[\sigma_{k,t}^2])^2} = \frac{1 - \frac{(\mathbf{E}[\sigma_{k,t}^2])^2}{\mathbf{E}[\sigma_{k,t}^4]}}{\kappa - \frac{(\mathbf{E}[\sigma_{k,t}^2])^2}{\mathbf{E}[\sigma_{k,t}^4]}} < \frac{1}{\kappa}. \end{aligned} \tag{18}$$

In the first line, we use that the variance of $\eta_{k,t+1}$ equals the expected squared error loss given in eq. (12). The first inequality in the second line is due to the insight that the expected squared error loss based on the noisy proxy is at least $(\kappa-1)\mathbf{E}[\sigma_{k,t}^4]$. The upper bound for R_k^2 that is given by equation (18) nicely illustrates that a low R_k^2 is not necessarily evidence for model misspecification but can simply be due to using a noisy volatility proxy. This point has been made before by Andersen and Bollerslev (1998), but for the special case of a one-step ahead forecast from a GARCH(1,1).[8] Note that the result in eq. (18) does not depend on the two-component structure of the model but is true for any conditionally heteroscedastic process.

Next, we derive an explicit expression for the Mincer-Zarnowitz R_k^2 of the M-GARCH model.

Proposition 3. *If $\sigma_{k,t+1}^2$ follows an M-GARCH process, Assumptions 1, 2 and 3 are satisfied and*

[8] See Andersen et al. (2005) for a model-free adjustment procedure for the predictive R^2.

$h_{k,t+1|t} = \tau_{t+1} g_{k,t+1|t}$, *the population* R_k^2 *of the Mincer-Zarnowitz regression is given by*

$$R_k^2 = \frac{\mathbf{Var}(h_{k,t+1|t})}{\mathbf{Var}(\varepsilon_{k,t+1}^2)} = \frac{\mathbf{E}[g_{k,t+1|t}^2]\mathbf{E}[\tau_{t+1}^2] - \mathbf{E}[\tau_{t+1}]^2}{\mathbf{E}[g_{k,t+1}^2]\mathbf{E}[\tau_{t+1}^2]\kappa - \mathbf{E}[\tau_{t+1}]^2} \tag{19}$$

with $\mathbf{E}[g_{k,t+1}^2]$ *as given by eq. (4) and where we use that* $\mathbf{E}[g_{k,t+1|t}] = \mathbf{E}[g_{k,t+1}] = 1$. *We obtain the following two properties:*

1. R_k^2 *decreases monotonically with increasing forecast horizon* k *and, in the limit, converges to*

$$R_\infty^2 = \frac{\mathbf{E}[\tau_{t+1}^2] - \mathbf{E}[\tau_{t+1}]^2}{\mathbf{E}[g_{k,t+1}^2]\mathbf{E}[\tau_{t+1}^2]\kappa - \mathbf{E}[\tau_{t+1}]^2}, \tag{20}$$

   where we use that $\lim_{k\to\infty} \mathbf{E}[g_{k,t+1|t}^2] = 1$.[9]

2. *As* $\mathbf{E}[\tau_{\tau+1}^2]$ *increases to infinity,* R_k^2 *increases monotonically to* $\mathbf{E}[g_{k,t+1|t}^2]/(\mathbf{E}[g_{k,t+1}^2]\kappa) \leq 1/\kappa$ *for all* k.

The first property rests on the insight that the forecast of the GARCH component converges to one (as $k \to \infty$) and, hence, the Mincer-Zarnowitz regression reduces to a regression of $\varepsilon_{k,t+1}^2$ on a constant and τ_{t+1}. Thus, the R_∞^2 can be interpreted as the fraction of the total variation in daily returns that can be attributed to the variation in the long-term component. This measure is directly related to the variance ratio statistic discussed in Section 2.2, eq. (7), which we will use for measuring the relative importance of the long-term component.

Second, the result that R_k^2 increases when τ_{t+1} gets more volatile may be puzzling at first but is due to the fact that the variance of $\varepsilon_{k,t+1}^2$ increases more than the variance of $h_{k,t+1|t}$. Again, this implies that for the very same model the R_k^2 will be higher in high volatility regimes than in low volatility regimes, which can be misleading when calculating the R_k^2 for different subsamples.

For illustration purpose, we consider a one-step ahead forecast and present R_1^2 directly as a function of the model parameters:

Lemma 1. *If* $\sigma_{k,t+1}^2$ *follows an M-GARCH, Assumptions 1, 2 and 3 are satisfied and* $h_{1,t+1|t} = \tau_{t+1} g_{1,t+1}$, *the population* R_1^2 *of the Mincer-Zarnowitz regression is given by*

$$R_1^2 = \frac{(1-(\alpha+\gamma/2+\beta)^2)\mathbf{E}[\tau_{t+1}^2] - (1-(\alpha+\gamma/2)^2\kappa - 2(\alpha+\gamma/2)\beta - \beta^2)\mathbf{E}[\tau_{t+1}]^2}{(1-(\alpha+\gamma/2+\beta)^2)\mathbf{E}[\tau_{\tau+1}^2]\kappa - (1-(\alpha+\gamma/2)^2\kappa - 2(\alpha+\gamma/2)\beta - \beta^2)\mathbf{E}[\tau_{t+1}]^2}. \tag{21}$$

Moreover, the limit of R_1^2 *in* $\mathbf{E}[\tau_{\tau+1}^2]$ *is*

$$\lim_{\mathbf{E}[\tau_{\tau+1}^2]\to\infty} R_1^2 = 1/\kappa.$$

For $\tau_{\tau+1}$ being constant and $\gamma = 0$, eq. (21) reduces to the expression in Andersen and Bollerslev (1998, p. 892) for the symmetric GARCH(1,1), i.e. $R_1^2 = \alpha^2/(1-2\alpha\beta-\beta^2)$. Lemma 1 shows that R_1^2

[9] Although by assumption $k \leq I_t$ in our setting, we can think of, for example, a biannual period and daily volatility forecasts. In this case k can be at most 132 $(= 6 \cdot 22)$. For such a large k and under reasonable assumptions on the GARCH parameters, we have $\mathbf{E}[g_{132,t+1|t}^2] \approx 1$.

reaches the upper bound when τ_{t+1} is getting more volatile. When τ_{t+1} gets more volatile, we find that the expected SE, $\mathbf{E}[\mathrm{SE}(\varepsilon_{1,t+1}^2, \tau_{t+1}g_{1,t+1})] = (\kappa-1)\mathbf{E}[\tau_{t+1}^2]\mathbf{E}[g_{1,t+1}^2]$ is increasing. Although this observation is correct, the variance of the squared returns, $\mathbf{Var}(\varepsilon_{1,t+1}^2) = \kappa\mathbf{E}[g_{1,t+1}^2]\mathbf{E}[\tau_{t+1}^2] - \mathbf{E}[\tau_{t+1}]^2$, is increasing even faster due to the constant κ compared to $\kappa - 1$, which leads to an overall increase in the R_1^2.

The effect of the long-term component on R_1^2 is depicted in Figure 2 which shows the population R_1^2 as a function of $\mathbf{E}[\tau_{t+1}^2]$ for different values of α and β. For τ_{t+1} being constant, the M-GARCH model reduces to a GARCH(1,1). As can be seen, even a small increase in $\mathbf{E}[\tau_{t+1}^2]$ can cause a steep increase in the population R_1^2. Also, for a given level of $\mathbf{E}[\tau_{t+1}^2]$, the population R_1^2 is the higher the more persistent (as measured by $\alpha + \gamma/2 + \beta$) the GARCH component is.

[Figure 2 here]

Finally, we consider cumulative volatility forecasts. The Mincer-Zarnowitz regression for evaluating the cumulative k-day volatility forecast is given by

$$\widetilde{RV}_{1:k,t+1} = \tilde{\delta}_0 + \tilde{\delta}_1 h_{1:k,t+1|t} + \eta_{1:k,t+1},$$

where the latent variance is replaced by the realized variance $\widetilde{RV}_{1:k,t+1} = \sum_{i=1}^k \varepsilon_{i,t+1}^2$ (purely based on daily data). The corresponding $R_{1:k}^2$ is given by

$$R_{1:k}^2 = \frac{\mathbf{Var}(h_{1:k,t+1|t})}{\mathbf{Var}(\widetilde{RV}_{1:k,t+1})} = \frac{\mathbf{E}[\tau_{t+1}^2]\mathbf{E}[(\sum_{i=1}^k g_{i,t+1|t})^2] - k^2\mathbf{E}[\tau_{t+1}]^2}{\mathbf{E}[\tau_{t+1}^2]\mathbf{E}[(\sum_{i=1}^k g_{i,t}Z_{i,t}^2)^2] - k^2\mathbf{E}[\tau_{t+1}]^2}. \tag{22}$$

As before, one can show that $R_{1:k}^2$ increases monotonically in $\mathbf{E}[\tau_{t+1}^2]$.

3 Simulation

Next, we turn to a Monte-Carlo simulation for evaluating the forecast performance of the M-GARCH model and to empirically illustrate our theoretical results. In addition, we examine what kind of losses in forecast performance can be expected when the true data generating process is indeed a M-GARCH model but misspecified models (in terms of lag length), a simple one-component GARCH model or the HAR model of Corsi (2009) are employed instead.

3.1 Data generating process

We simulate an intra-day version of the two component GARCH model as

$$\varepsilon_{n,i,t} = \sqrt{g_{i,t}\tau_t}\frac{Z_{n,i,t}}{\sqrt{N}}, \tag{23}$$

where the index $n = 1, \ldots, N$ denotes the intra-day frequency and $Z_{n,i,t} \overset{iid}{\sim} \mathcal{N}(0,1)$. We generate $N = 288$ intra-day returns. Note that $g_{i,t}$ still fluctuates at the daily and τ_t at the low-frequency. Hence, by aggregating returns to a daily frequency $\varepsilon_{i,t} = \sum_{n=1}^{N} \varepsilon_{n,i,t}$ the model in eq. (23) is consistent with our daily model. Simulating intra-day data allows us to calculate the daily realized variance, $RV_{i,t} = \sum_{n=1}^{N} \varepsilon_{n,i,t}^2$, as a precise measure of the daily conditional variance. Similarly, we obtain the realized variance over the first k days of month t as $RV_{1:k,t} = \sum_{i=1}^{k} RV_{i,t}$. We simulate data for a period of 40 years of intra-daily returns (from which we construct 10560 daily return and realized variance observations). For each model specification, we perform 2000 Monte-Carlo replications. The alternative model specifications are described in the following:

Specification 1: The first specification assumes that τ_t fluctuates at a monthly frequency. We assume that each month consists of $I_t = 22$ days. Thus, we consider a truly mixed-frequency setting. The long-term component is defined as in eq. (5) and (6) with $m = 0.1$, $\theta = 0.3$, $w_1 = 1$, $w_2 = 5$ and $K = 36$. The choice of three years as MIDAS lag length follows Conrad and Loch (2015). Setting $w_2 = 5$ implies a monotonically decaying weighting scheme with weights close to zero for lags greater than two-thirds of K. The explanatory variable X_t is assumed to follow an AR(1) process, $X_t = \phi X_{t-1} + \xi_t$, $\xi_t \overset{i.i.d.}{\sim} \mathcal{N}(0,\ \sigma_\xi^2)$, with $\psi = 0.95$ and $\sigma_\xi^2 = 0.2^2$. Using these parameter values leads to a variance ratio (VR) of 20.40% when averaged over the 2000 Monte-Carlo simulations (recall that the VR was defined in eq. (7)). This size of the VR is comparable to the VRs typically found im empirical applications in mixed-frequency settings with daily returns and monthly macroeconomic variables (see, e.g., Conrad and Loch, 2015).

Specification 2: The second specification assumes that both components fluctuate at a daily frequency (i.e. $I_t = 1$). The parameters of the long-term component are chosen as $m = -0.1$, $\theta = 0.3$ with $K = 264$. The weighting parameters $w_1 = 1$ and $w_2 = 5$ are chosen as above. Choosing a lag length of roughly one year is motivated by our empirical results in Section 4 when estimating an M-GARCH model using $RV_{i,t}$ as the explanatory variable. In addition, we choose $\psi = 0.98$ and either $\sigma_\xi^2 = 0.2^2$ or $\sigma_\xi^2 = 0.5^2$. In our simulations, the former choice leads to an average variance ratio of 32.50% (low VR) and the latter to a variance ratio of 74.80% (high VR).

In both cases, the parameters of the GARCH-component, $g_{i,t}$, are given by $\alpha = 0.06$, $\beta = 0.91$ and $\gamma = 0$.

3.2 Parameter estimates

We use the first twenty years of the simulated data to produce in-sample estimates of the M-GARCH model. For each specification, we estimate three variants of the M-GARCH model. While in all three variants the structure of the short- and long-term component is correctly specified, only in one specification the lag length K corresponds to the true one while in the other two specifications K is chosen too small.[10] In addition, we estimate a simple (but misspecified) GARCH model that is obtained when restricting τ_t to be constant. The M-GARCH models as well as the simple GARCH are estimated by quasi-maximum likelihood (QML).

Table 1 reports the median and the interquartile range of the estimates for each parameter across the 2000 Monte-Carlo simulations. We observe that the median parameter estimates for the correctly specified M-GARCH models with $K = 36$ (monthly) and $K = 264$ (daily) are very close to the true parameter values. Choosing a lag length that is way too small $K = 12$ (monthly) or $K = 66$ (daily) leads to an estimate of θ which is considerably smaller than the true value. At the same time, the M-GARCH models with misspecified lag length lead to estimates of w_2 that are considerably smaller than $w_2 = 5$. By doing so, the model estimates try to offset the effect of misspecifying K. Figure 3 shows the estimated weighting schemes. While a modest misspecification of K still leads to an acceptable weighting scheme, the model with $K = 12$ (monthly) and the model with $K = 66$ (daily) lead to severely biased weighting schemes.

[Table 1 here]

Table 1 also shows that the GARCH parameter estimates are close to the true parameter values for the specifications with monthly τ_t and daily τ_t with low VR, but not for the one with daily τ_t and high VR. For this latter specification, the simple GARCH is severely misspecified and the estimate of α is considerable above 0.06. For this model the estimated persistence as measured by $\hat{\alpha} + \hat{\beta}$ is above 0.99 so that the estimated parameters clearly suffer from the so-called IGARCH effect (see Baillie et al., 1996, or Hillebrand, 2005).[11]

[Figure 3 here]

The second last column shows the average excess kurtosis of the fitted standardized residuals. Those residuals are given by $\varepsilon_{i,t}/\sqrt{\hat{\tau}_t \hat{g}_{i,t}}$ for the M-GARCH models and by $\varepsilon_{i,t}/\sqrt{\hat{g}_{i,t}}$ for the GARCH

[10] We do not report results for K being chosen too large as the beta-weighting scheme is flexible enough to set uninformative lags to almost zero. For an illustration see Figure 3.

[11] Recall from eq. (8) that the M-GARCH can be rewritten as a GARCH with time-varying intercept. For this type of process, Hillebrand (2005) shows that the sum of the estimated α and β parameters is biased towards one when the existing parameter change is not accounted for.

model. The only model for which the interquartile range does not cover the zero, is the GARCH model in the scenario with daily τ_t and high VR. This confirms our conjecture from Section 2.3 that the deGARCHed residuals from a misspecified GARCH model may still exhibit excess kurtosis if the true data generating process is an M-GARCH and the long-term component is sufficiently relevant.

Finally, the last column shows the average VRs that are implied by the estimated parameter values. Clearly, the estimated VRs are close to the actual ones when the lag order is correctly chosen, but are too low for choices of K that are too small (i.e. when $K = 12$ for monthly τ_t and $K = 66$ for daily τ_t).

3.3 Forecast evaluation

Next, we use the in-sample parameter estimates from the previous section and construct out-of-sample volatility forecasts for the remaining twenty years to evaluate the forecast performance of the different specifications. Recall that in our simulation, each month consists of 22 trading days. For each model, we construct k-step ahead volatility forecasts as well as the corresponding k-period cumulative volatility forecasts with $k = 1, \ldots, 22$. Based on the information available at the last day of the current month, forecasts are constructed for the following month. Thus, the simulation scenario with a monthly τ_t is perfectly in line the setting introduced in Section 2.5. In the simulation variant with daily τ_t, the optimal forecast would require a prediction of the long-term component. Instead, we simply fix the long-term component at its current level. This approach is commonly taken in the literature when the forecast horizon is beyond I_t.[12] Note that by construction we obtain only non-overlapping volatility forecasts. As before, the forecast performance is evaluated over the 2000 Monte-Carlo replications.

For each model, we present the average SE and the average QLIKE for the k-period cumulative volatility forecast as well as the R^2 of the Mincer-Zarnowitz regression. When evaluating the volatility forecasts, we employ $RV_{1:k,t}$ as our proxy for the latent volatility.[13]

In addition to the M-GARCH models and the GARCH model, we also consider a HAR specification that directly models the realized variances (see Corsi, 2009). We specify the HAR model in terms of the log of the realized variances. Using the notation $RV_s = RV_{i,t}$ with $s = s(i,t) =$

[12] Holding τ_t constant when forecasting is reasonable if τ_t changes smoothly and the forecast horizon is not "too large". Otherwise, one may use predictions of X_t, e.g., survey or time series forecasts, for calculating predictions of τ_t (see Conrad and Loch, 2015).

[13] Results based on the squared returns as the volatility proxy are available upon request.

$22 \cdot (t-1) + i$ the model for forecasting the k-period cumulative variance is given by

$$\log\left(\frac{RV_{s+1:s+k}}{k}\right) = b_0 + b_1 \log RV_s + b_2 \log\left(\frac{RV_{s-4:s}}{5}\right) + b_3 \log\left(\frac{RV_{s-21:s}}{22}\right) + \zeta_{s,k} \qquad (24)$$

with $RV_{s+1:s+k} = \sum_{i=1}^{k} RV_{s+i}$. The HAR model is estimated by OLS. Realized variance forecasts are obtained as follows:

$$RV_{s+1:s+k|s} = k \cdot \exp\left(b_0 + b_1 \log RV_s + b_2 \log\left(\frac{RV_{s-4:s}}{5}\right) + b_3 \log\left(\frac{RV_{s-21:s}}{22}\right) + \frac{1}{2}\mathbf{Var}(\zeta_{s,k})\right),$$

assuming the residuals $\zeta_{s,k}$ to be normally distributed. Note that there are two major differences between the forecasts from M-GARCH models and the HAR model. First, the HAR exploits the high-frequency information from the intra-day returns while the M-GARCH models are purely based on daily returns and the explanatory variable X_t. Second, the volatility forecasts of the HAR model are tailored to the forecast horizon (which requires the estimation of the HAR model for each forecast horizon) and, hence, are *direct* forecasts, while the M-GARCH forecasts are *iterated*.

3.3.1 Squared error loss

The upper four panels of Figure 4 and Figure 5 show the average SE for the cumulative k-step forecasts for the monthly- and daily-varying long-term component with a high variance ratio.[14] The upper left panel shows the average SE for the full out-of-sample period. In the three remaining panels, we distinguish between three different volatility regimes: *low*, *normal* and *high*. An observation is falling in the low/normal/high volatility regime, if the level of volatility on the day the forecast has been issued is below the 25%-quantile, between the 25%- and 75%-quantile or above the 75%-quantile of the distribution of historical volatility.

In the full sample, all M-GARCH models outperform the HAR model. Interestingly, misspecifying K does hardly appear to affect forecast performance. While the GARCH model performs reasonably well for the monthly τ_t with low VR, it clearly produces the worst forecasts for the daily τ_t with high VR. In line with our theoretical results for the expected SE error, eq. (16), we observe substantially higher values for the SE loss in the high volatility regime compared to the SE loss in the low volatility regime. This holds for both the monthly and daily long-term components. Thus, our simulations suggest that the ability of tests for comparing the predictive accuracy of two competing models based on the SE loss difference may be sensitive with respect to the current volatility regime.

[14] Figure 15 in Appendix E displays the case of a daily-varying τ_t with a low variance ratio.

[Figure 4 & 5 here]

3.3.2 QLIKE loss

The lower panels of Figure 4 and Figure 5 show the average QLIKE for cumulative k-period forecasts. In sharp contrast to the evaluation based on the SE, it becomes evident that the QLIKE loss is robust with respect to the volatility regime, i.e. for each model the size of the loss is roughly the same across the three volatility regimes. We further investigate this issue by conducting Diebold and Mariano (1995) tests on the loss differences between the correctly specified M-GARCH and the HAR model for forecast horizons of $k \in \{1, 5, 10, 20\}$ days. We focus on the case of a daily τ_t in combination with a high VR. For each of the 2000 Monte-Carlo simulations, we perform a Diebold-Mariano test based on the full sample as well as on the low/normal/high volatility regime. In Table 7, Panel A summarizes the results for the Diebold-Mariano tests when the SE loss is employed and Panel B when the QLIKE loss is employed.[15] The table reports the empirical rejection rates of the Diebold-Mariano test at the 5% nominal level. For example, the number 0.814 indicates that in 81.4% of the 2000 tests based on the SE and the full sample, the null hypothesis of equal forecast accuracy was rejected in favor of the M-GARCH model.[16] As expected, the Diebold-Mariano tests are most powerful for $k = 1$ and have decreasing power with increasing forecast horizon. Also, note that among the three volatility regimes for both loss functions the power of the Diebold-Mariano test is the highest in the normal regime.[17] Most importantly, the Diebold-Mariano tests based on the QLIKE in Panel B are always more powerful than tests based on the SE. This effect is the strongest in the high volatility regime. For example, for $k = 1$ the difference in the empirical power is 20.2 percentage points in the high volatility regime. This difference in the power stays roughly constant when increasing the forecast horizon. Based on these power comparisons, again the QLIKE loss appears to be the preferred choice.

[Table 2 here]

3.3.3 Mincer-Zarnowitz regression

Similarly, Figure 6 shows the $R^2_{1:k}$ of the Mincer-Zarnowitz regressions for the k-period ahead forecasts for the full out-of-sample period as well as the three volatility regimes. Panel (a) shows the results for monthly τ_t and Panel (B) the results for daily τ_t with a high variance ratio.[18] The

[15] Equivalent DM test results for monthly τ_t and daily τ_t with low VR can be found in Appendix D, Tables 7 and 8.

[16] There is also a small number of rejections in favor of the HAR model. We do not report those numbers.

[17] However, the tests in the normal regime are – by construction – based on twice as many observations as in the low/high volatility regime, which may affect this finding.

[18] The case of a daily τ_t and a low VR can be found in Appendix E, Figure 16.

$R^2_{1:k}$'s for the M-GARCH models are clearly higher than for the HAR model and decrease for all models with the forecast horizon. In line with our theoretical considerations, among the three volatility regimes we observe the highest $R^2_{1:k}$'s in the high volatility regime. Clearly, these high $R^2_{1:k}$'s do not reflect improved forecast performance but are an artefact due to the construction of the $R^2_{1:k}$. Even more misleading, the $R^2_{1:k}$ in the full sample is higher than in each subsample.

[Figure 6 here]

In summary, the M-GARCH models clearly outperform the misspecified GARCH model and the HAR model in terms of forecast performance. That is, a direct modeling of the realized variance is no substitute for a proper identification of the underlying component structure. Moreover, in line with our considerations in Section 2.5, the simulation results suggest that the QLIKE should be the preferred loss function when comparing forecast performance of two competing volatility models.

4 Empirical Analysis

Last, we turn to an empirical application of the M-GARCH models to S&P 500 return data. Specifically, we will focus on model forecast performance conditional on the volatility regime in which the forecasts are produced. Making use of our flexible setup, we employ explanatory variables at different frequencies in a real-time forecast evaluation. In Section 4.1 we first introduce our data set and full-sample estimation results for various M-GARCH models are reported in Section 4.2. Thereafter, in Section 4.3 we explain how real-time volatility forecasts can be constructed when taking into account the release schedule of the macroeconomic explanatory variables. The forecast comparison is carried out in Section 4.4.

4.1 Data

Stock market data. We consider daily log-returns on the S&P 500, calculated as $r_{i,t} = 100 \cdot (\log(p_{i,t}) - \log(p_{i-1,t}))$, for the 1971:M1 to 2017:M2 period. For the evaluation of the volatility forecasts, we employ daily realized variances, $RV_{i,t}$, based on five-minute intra-day returns obtained from the Realized Library of the Oxford-Man Institute of Quantitative Finance (see Heber et al., 2009). Daily realized variances are available from the year 2000 onwards.

Explanatory variables. As explanatory variables in the τ_t component, we use either financial or macroeconomic variables. As financial variables we employ a rolling window of the average realized volatility during the past 22 days based on squared daily returns, $\sqrt{\widetilde{\text{RV}}_{i,t}(22)} = \sqrt{1/22 \sum_{j=0}^{21} r^2_{i-j,t}}$,

the VIX index (converted to a daily level), $\text{VIX}_{i,t}/\sqrt{252}$, as a forward-looking market-based measure of the next month's volatility and the NFCI. Being a measure for the tightness of financial conditions in the US, the NFCI is a weighted average of 105 standardized financial indicators of risk, credit and leverage derived by dynamic factor analysis. Constituents are, among others, loan spreads of financial institutions, equity and bond price measures, as well as figures derived from the Senior Loan Officer Opinion Survey on Bank Lending Practices.[19] As an example, weights for tightening standards on various loans being positive implies an increase in the NFCI whenever banks become more conservative in lending. The NFCI is released on a weekly frequency. The methodological construction of the NFCI is similar to the Aruoba-Diebold-Scotti Business Conditions Index (Arouba et al., 2009).

Macroeconomic conditions are measured by the Chicago Fed national activity index (NAI) and growth rates of industrial production and housing starts, both calculated as $\Delta X_t = 100 \cdot (\ln(X_t) - \ln(X_{t-1}))$. The three macroeconomic variables are available at a monthly frequency. More detailed information on the data set is provided in Appendix F.

Table 3 provides summary statistics for the stock returns as well as the six explanatory variables. While the macroeconomic variables are included from 1971 onwards, the NCFI series begins in 1973 and the VIX is available from 1990 onwards. In order to facilitate comparison between $\text{VIX}_{i,t}/\sqrt{252}$ and $\sqrt{\widetilde{\text{RV}}_{i,t}(22)}$, the summary statistics for the latter cover the 1990:M1 to 2017:M2 period. Figure 7 shows the evolution of the corresponding time series.

[Table 3 & Figure 7 here]

Before we estimate M-GARCH models, we employ the Conrad and Schienle (2018) Lagrange multiplier (LM) test for an omitted multiplicative component in simple one-component GARCH models. This test checks whether a simple GJR-GARCH(1,1) is misspecified in the sense of neglecting an omitted component which is driven by the explanatory variable X_t. Since the test is of the LM type, it requires the estimation of the model under the null hypothesis only. Assuming that under the alternative there is a second component which is driven by K lags of the variable X_t, the test statistic can be shown to be χ^2 with K degrees of freedom. An appealing property of the test is that it can be applied in settings where X_t is observed at the same frequency as the returns but also when X_t is observed at a lower frequency. Intuitively, the test checks whether the squared standardized residuals from the GJR-GARCH that is estimated under H_0 are predictable using (functions of) past values of X_t. Table 4 shows the outcome of the test when applied to each of our explanatory variables. When either choosing $K = 1$ or $K = 2$, the test results clearly suggest

[19] https://www.federalreserve.gov/data/sloos/sloos.htm

that for all variables but housing starts we can reject the null hypothesis that a GJR-GARCH is correctly specified. Thus, the LM test results justify the usage of the M-GARCH models using our X_t variables. Estimating an M-GARCH model for housing starts will also reveal why the test does not reject for this variable when $K = 1$ or $K = 2$.

[Table 4 here]

4.2 Full sample parameter estimates

We first estimate an M-GARCH model for each explanatory variable for the full sample, whereby we include a constant in the mean equation, i.e. we model returns as $r_{i,t} = \mu + \varepsilon_{i,t}$. The corresponding QML estimates are reported in Table 5. For the monthly macroeconomic variables, we follow Conrad and Loch (2015) and choose $K = 36$. Based on the Bayesian information criterion (BIC) and visual inspection of the estimated weighting schemes for alternative choices of K, we select $K = 52$ for the NFCI, $K = 3$ for the VIX and $K = 264$ for $\sqrt{\widetilde{\text{RV}}_{i,t}(22)}$. Thus, for the forward-looking VIX only the most recent information appears to drive long-term volatility, while the backward-looking $\sqrt{\widetilde{\text{RV}}_{i,t}(22)}$ is smoothed over many lags.[20] For all variables but housing starts, we find that a restricted Beta weighting scheme with $w_1 = 1$ is the best choice, i.e. the optimal weights are declining from the beginning. For housing starts, an unrestricted weighting scheme which allows for 'hump-shaped' weights is required. This confirms the finding in Conrad and Loch (2015) that housing starts is leading with respect to long-term volatility.[21] Figure 8 shows the estimated weighting schemes. Note that the M-GARCH models based on the NFCI and the three macroeconomic variables employ return data for the 1974:M1 to 2017:M2 period, while the models with daily τ_t employ data for 1990:M1 to 2017:M2 only. Hence, it is not possible to compare the log-likelihood function (LLF) and the BIC for models based on daily τ_t with models based on weekly/monthly τ_t.

[Table 5 here]

Concerning the QML estimates, it is interesting to observe that the M-GARCH models with daily τ_t lead to lower estimates of β than the models with weekly/monthly τ_t. Also, the models with daily τ_t are characterized by estimates of α that are basically zero. These parameter estimates imply that the deviations of the short-term component from the long-term component are more short-lived for the M-GARCH models with daily τ_t. This behavior is also evident from Figure 9 which shows the evolution of the annualized long- and short-term components.[22] The signs of the estimated θ's

[20] A more detailed comparison for different values of K is available upon request.

[21] Because of this leading property the LM test did not reject the null hypothesis for $K = 1$ or $K = 2$. However, choosing $K = 10$ leads to a rejection of the null.

[22] To ensure comparability across the six M-GARCH models, all figures start in 2000.

for realized volatility, the VIX and the macroeconomic variables are in line with the findings in the previous literature. Higher levels of financial volatility tend to increase long-term volatility, while an improvement of macroeconomic conditions decreases long-term volatility. This counter-cyclical behavior of financial volatility has been previously documented. The finding that tighter financial conditions (i.e. an increase in the NCFI) predict higher volatility is new. While the positive relation between realized volatility and long-term volatility might be viewed as 'mechanical', the NCFI as well as the macroeconomic variables can be considered as fundamental drivers of financial volatility.

[Figure 8 & 9 here]

Again, we gauge the importance of the variation in the long-term component for the overall expected variation in return volatility by the variance ratio introduced in eq. (7). To facilitate comparison across models, we focus on the monthly variation of volatility. That is, for all models we denote by σ_M^X the monthly aggregate volatility. For models with monthly long-term component, we have that $\tau_M^X = \tau_t^X$. For models with daily or weekly long-term component, τ_M^X refers to monthly aggregates of the daily/weekly τ_t^X. We then calculate $\text{VR}(X) = \mathbf{Var}(\log(\tau_M^X)/\mathbf{Var}(\log(\sigma_M^X))$, where the X indicates that the variance ratio is based on a specific explanatory variable. As Table 5 shows, the models with daily τ_t achieve much higher variance ratios than the models with weekly/monthly τ_t. For example, the variance ratio of 75.31% for the model based on the VIX implies that two-thirds of the expected variation in return volatility can be traced back to variation in the VIX. Naturally, the models based on monthly macroeconomic variables attain lower variance ratios of about 15%. This value is comparable to the figures reported in Conrad and Loch (2015) for quarterly and monthly data.

4.3 Real-time estimates

In the following, we make use of vintage data. This allows for a realistic evaluation of the GARCH-MIDAS models' ability to describe the behavior of long-term financial volatility in real time. For a comparison of the full sample estimate of the long-term component with the real-time estimates, we reestimate all six GARCH-MIDAS models presented in Table 5 on a monthly basis. Estimation is performed on a rolling window of observations whereby the window size is chosen as the length of the first estimation period ending in 2010:M12. The period 2010:M01 to 2017:M2 will also be used as the out-of-sample period for the forecast evaluation in Section 4.4. In order to ensure that our estimates of the long-term component are feasible in real-time, we employ vintage data that is available for the NFCI, the NAI, IP and housing starts from the ALFRED database hosted by the

St. Louis Fed.[23] For the three financial variables, parameter reestimation is carried out at the end of each month. When employing macroeconomic variables, we reestimate the models on the days on which new releases of X_t become available.

For example, the real-time figures of industrial production for the current month are unavailable until (roughly) the second week of the following month. This type of release schedule is illustrated in Figure 10. In theory, in our model the value of the long-term component in March, τ_{Mar}, depends on the explanatory variable of February, X_{Feb}, and previous months. However, in the beginning of March, X_{Feb} is not available until the midst of the month, e.g. March 15. Hence, between March 1 and 14, we are unable to calculate τ_{Mar} and fix the value of the long-term component at $\hat{\tau}_{March} = \tau_{Feb}$ instead. Once X_{Feb} is released, we update τ_{Mar} by incorporating X_{Feb}.

[Figure 10 here]

Figure 11 shows the estimated long-term components based on the full-sample estimates (as reported in Table 5, dotted lines) and based on the rolling window real-time estimates (solid lines). For $\sqrt{\widetilde{\text{RV}}(22)}$ as well as $\text{VIX}/\sqrt{252}$ the full-sample and rolling window long-term components might differ, because they are based on distinct sample periods. For the NFCI, the NAI, IP and housing starts, the two long-term components are not only based on distinct sample periods but also on different real-time vintages.[24] Although Figure 11 might suggest that there are only mild differences between the full-sample and real-time estimates of the long-term component, the average absolute differences are quite sizable. For example, the average absolute difference between the full-sample and real-time estimates based on industrial production is 8.28% which is huge compared to the 2.18% mentioned in footnote 24. Similar numbers are obtained for the other variables: 6.94% for housing starts, 3.6% for the NAI, 5.18% for the NFCI, 5.18% for the VIX and 9.98% for realized volatility. In summary, these figures highlight to importance of using real-time vintage data instead of final data releases for a meaningful evaluation of the M-GARCH models forecast accuracy.

[Figure 11 here]

4.4 Forecast evaluation

We now evaluate the forecasting performance of the different M-GARCH models out-of-sample. Volatility forecasts are based on the models that have been estimated in real-time. In order to generate a sufficient number of out-of-sample predictions, we construct cumulative volatility forecasts

[23] For more details on real-time availability see Appendix F.

[24] For example, according to Croushore (2011) the mean absolute revision from the initial release to the latest available data was 2.18% for industrial production during the 1965:Q3 to 2006:Q4 period. Among the variables considered in Croushore (2011) this is the highest value (even higher than for GDP).

for the next 22 trading days on a daily basis (and not only at the beginning of each month). When calculating the forecasts, we always keep the long-term component fixed at its current level.

As natural competitors, we consider the nested GARCH model as well as the HAR model described in eq. (24). Figure 12 shows the one-day ahead volatility forecasts of the HAR model based on the full sample along with realized volatilities (left panel) as well as the parameter estimates for b_1, b_2 and b_3 when the forecast horizon is either 1-, 5- or 22-days (right panel). As the right panel shows, most weight is given to the most recent daily realized volatility when the forecast horizon is one day. On the other hand, for a forecasting horizon of 22 days the highest weight is attached to the monthly realized volatility. For the forecast evaluation, both the GARCH and the HAR model are reestimated on a monthly basis using a rolling window of past observations. Finally, we consider a no-change volatility forecast which simply scales today's realized variance to the appropriate horizon: $h_{s+1:s+k|s} = k \cdot RV_s$.

[Figure 12 here]

We provide a comparison of the forecast performance of the different models based on the SE loss function, SE $\left(RV_{s+1:s+k}, \hat{h}^{(j)}_{s+1:s+k|s}\right)$, as well as the QLIKE loss, QLIKE $\left(RV_{s+1:s+k}, \hat{h}^{(j)}_{s+1:s+k|s}\right)$, where $RV_{s+1:s+k}$ denotes the realized variance over the $s+1$ to $s+k$ period and $\hat{h}^{(j)}_{s+1:s+k|s}$ is the corresponding forecast based on model j. In addition to the discussion in Section 2.5, it is important to highlight that the QLIKE loss is an asymmetric loss function in the sense that it punishes underestimation more heavily than overestimation, while the SE loss function is symmetric. A preliminary check of the forecasts from the different models suggests that this distinction might be important. Table 6 shows the average absolute forecast error as well as the average standardized forecast error for all models and horizons of 1-, 8- and 15-days / 1-, 2- and 3-months ahead. While the HAR appears to generate forecast errors that are the smallest in absolute value, the average standardized forecast error is smaller than one for all M-GARCH models but bigger than one for the HAR model. Thus, we might expect that the SE should favor the HAR model but the QLIKE an M-GARCH specification.

[Table 6 here]

For testing whether there is one or several models that significantly outperform the others, we follow the Model Confidence Set (MCS) approach introduced by Hansen et al. (2011). We denote by $\mathcal{M}$ the set of all competing models, i.e. the various M-GARCH models, the HAR model, as well as the simple GARCH and the no-change model. We define $d_{s,i,j} = L(RV_{s+1:s+k}, \hat{h}^{(i)}_{s+1:s+k|s}) - L(RV_{s+1:s+k}, \hat{h}^{(j)}_{s+1:s+k|s})$ as the difference in the loss of model i and model j for the forecast based

on period s. The loss function L is either the SE or the QLIKE loss. Next, we construct the average loss difference, $\bar{d}_{i,j}$, and calculate the test statistic

$$t_{ij} = \frac{\bar{d}_{i,j}}{\sqrt{\widehat{\mathbf{Var}}\left(\bar{d}_{i,j}\right)}} \text{ for all } i, j \in \mathcal{M}. \tag{25}$$

The MCS test statistic is then given by $T_{\mathcal{M}} = \max_{i,j \in \mathcal{M}} |t_{i,j}|$ and has the null hypothesis that all models have the same expected loss. Under the alternative, there is some model i that has an expected loss that exceeds the expected loss of all other models $j \in \mathcal{M} \setminus i$. If the null is rejected, the worst performing model is eliminated. The test is performed iteratively, until no further model can be eliminated. We denote the final set of surviving models by $\mathcal{M}_{MCS}$. This final set contains the best forecasting model with confidence level $1 - \nu$. In the empirical application, we set $\nu = 0.1$.

Since the asymptotic distribution of the test statistic $T_{\mathcal{M}}$ is nonstandard, we approximate it by block-bootstrapping as proposed by Hansen et al. (2011), where the block length is determined by fitting an AR(p) process on the series of loss differences. In our analysis, 40.000 bootstrap replications at each stage were sufficient in order to obtain stable results. For implementing the MCS procedure, we use the R-package *MCS* (see Bernardi and Catania, 2015).

4.4.1 Forecasting cumulative 1- to 22-days ahead volatility

Figure 13 presents the outcome of the MCS procedure when applied to cumulative 1- to 22-days ahead volatility forecasts. In each panel the solid lines correspond to the different models, whereby the lowest line corresponds to the best model (rank 1) and the highest line to the worst model (rank 9). For a given forecast horizon, all models that are included in the final set, $\mathcal{M}_{MCS}$, are indicated by a dot. For both loss functions, we provide a forecast evaluation based on the full out-of-sample period, but also on subsamples of low, normal and high volatility. We define these regimes based on the level of realized volatility on the day at which the forecast is issued. That is, a forecast belongs to a low, normal or high volatility regime if the realized volatility on the day the forecast is produced was below its historical 25% quantile, in between the 25% and 75% quantile or above its 75% quantile. In total, we have 599 observations in the low, 868 in the normal, and 267 in the high regime.

[Figure 13 here]

We begin with discussing the results for our favorite measure of forecast performance. The upper four panels of Figure 13 show the forecast evaluation based on the QLIKE loss. First, note that in the low volatility regime, the model based on the VIX and the HAR are included in the MCS,

whereby the VIX-based model is almost always ranked first. The same two models are the only models that are included in the MCS for all forecast horizons during the normal volatility regime. In the high volatility regime the QLIKE essentially is not able to distinguish between the different models, so that basically all of them are included in the model confidence set. The latter finding may be driven by the fact that forecast performance of all models substantially deteriorates during the high volatility regime such that distinguishing between models is impossible. Finally, the full sample results for the QLIKE appear to be driven by the low/normal volatility level and clearly favor the VIX-based M-GARCH model and, to a lesser extent, the HAR.

The lower four panels of Figure 13 show the forecast evaluation based on the SE loss. While for the low and normal volatility regime the VIX-based M-GARCH as well as the HAR are almost always included in the MCS, the HAR model appears to significantly outperform all other models in the high volatility regime (for essentially all horizons).[25] In contrast to our finding for the QLIKE, it appears that the MCS based on the SE for the full sample is entirely driven by the results for the high-volatility regime. However, the strong evidence in favor of the HAR in the high volatility regime is driven by a few extreme observations. This again questions the results based on the SE. Thus, we consider the QLIKE-based results as being more credible.

4.4.2 Forecasting volatility 2- and 3-months ahead.

Finally, we forecast volatility for 1-, 2-, and 3-months ahead, i.e. 1–22, 23–44 and 45–66-days ahead. Figure 14 shows the corresponding MCS rankings. Note that the ranking for one-month ahead is exactly the same as the one obtained for 22-days in Figure 13 and is only included to simplify the comparison across horizons. In the full sample, the rankings based on the QLIKE loss function favor the model based on housing starts at the 2- und 3-months horizon. For 3-months ahead, die QLIKE favors the VIX-based model in the low volatility regime and the NFCI-based model in the high-volatility regime. Based on the SE and for the full sample, the best model in the MCS for 2- and 3-months ahead is the M-GARCH based on the NCFI. Again, the full sample outcome for the SE appears to be driven by the high volatility sample which produces a similar ranking.

In summary, when forecasting volatility 2- or 3-months ahead and based on the model rankings, the M-GARCH models based on the NCFI or the VIX improve upon the HAR model independently of the loss function under consideration. Moreover, the M-GARCH model based on housing starts is the only model based on a macroeconomic variable that is repeatedly included in the MCS. Thus, our empirical results confirm previous findings that macroeconomic and financial conditions are

[25] At least for some periods, the no-change forecast is included in the MCS of the normal regime. The relatively good performance of the no-change forecast in the normal regime might be explained by the mean-reverting behavior of volatility. If volatility is at its long-term average, the no-change forecast is a good prediction.

useful predictors for longer-term financial volatility (see, for example, Christiansen et al., 2012, and Paye, 2012).

[Figure 14 here]

5 Conclusions

We derive and discuss the properties of the kurtosis of the returns and the ACF of the squared returns generated by a multiplicative two-component GARCH model. It is shown that a multiplicative GARCH model leads to returns with higher kurtosis than the nested one-component GARCH specification and that the ACF of the squared returns can be much more persistent than in the nested GARCH model. If the long-term component is sufficiently persistent, the ACF shows long-memory type behavior.

In addition, we show that the SE loss has several shortcomings when used to evaluate the precision of volatility forecasts generated from a mixed-frequency two-component model. In particular, the expected SE loss will vary with the volatility regime. When using the R^2 of a Mincer-Zarnowitz regression as a measure of the goodness of the forecast, this leads to the unpleasant property that the R^2 will be the highest in the regime with the highest volatility. In contrast, the QLIKE loss is shown to be independent of the volatility regime and, hence, is our preferred measure of forecast accuracy.

By means of a Monte-Carlo simulation we illustrate that the QLIKE is indeed better suited for distinguishing between competing models. In particular, we show that Diebold-Mariano tests based on the QLIKE are more powerful than tests based on the SE loss. These findings complement the arguments in Patton and Sheppard (2009) and Patton (2011).

In an empirical application to S&P 500 stock returns and U.S. macroeconomic and financial data, we compare the forecast performance of the M-GARCH models with the popular HAR model for several forecast horizons. To allow for a realistic forecast comparison, we estimate all models based on a rolling window of observations and employ real-time vintage data. Following Hansen (2011), we base the model comparison on model confidence sets. Under the QLIKE loss and for short-term forecasts (up to one month), an M-GARCH model based on the VIX and the HAR model are the preferred specifications according to the MCS. When forecasting 2- to 3-months ahead, lower-frequency financial and macroeconomic variables (housing starts and the NFCI) are the most relevant predictors of financial volatility. For these forecast horizons and based on the model rankings, the M-GARCH models outperform the HAR model

References

Amado, C., Silvennoinen, A., Teräsvirta, T., 2018. Models with Multiplicative Decomposition of Conditional Variances and Correlations. Discussion Paper.

Amado, C., Teräsvirta, T., 2013. Modelling Volatility by Variance Decomposition. *Journal of Econometrics*, 175, 142-153.

Amado, C., Teräsvirta, T., 2017. Specification and Testing of Multiplicative Time-Varying GARCH Models with Applications. *Econometric Reviews*, 36, 421-446.

Amendola, A., Candila, V., Scognamillo, A., 2017. On the Influence of US Monetary Policy on Crude Oil Price Volatility. *Empirical Economics*, 52, 155-178.

Andersen, T. G., Bollerslev, T., 1998. Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts. *International Economic Review*, 39, 885-905.

Andersen, T. G., Bollerslev, T., Meddahi, N., 2005. Correcting the Errors: Volatility Forecast Evaluation Using High-Frequency Data and Realized Volatilities. *Econometrica*, 73, 279-296.

Asgharian, H., Hou, A. J., Javed, F., 2013. The Importance of the Macroeconomic Variables in Forecasting Stock Return Variance: A GARCH-MIDAS Approach. *Journal of Forecasting*, 32, 600-612.

Aruoba, S. B., Diebold, F. X., Scotti, C., 2009. Real-Time Measurement of Business Conditions. *Journal of Business and Economic Statistics*, 27, 417-427.

Baillie, R. T., Bollerslev, B., 1992. Prediction in Dynamic Models With Time-Dependent Conditional Variances. *Journal of Econometrics*, 52, 91-113.

Baillie, R. T., Bollerslev, T., Mikkelsen, H. O., 1996. Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, 74, 3-30.

Baillie, R. T., Morana, C., 2009. Modelling Long Memory and Structural Breaks in Conditional Variances: An Adaptive FIGARCH Approach. *Journal of Economic Dynamic and Control*, 33, 1577-1592.

Bernardi, M., Catania, L., 2015. The Model Confidence Set Package for R. CEIS Working Paper No. 362.

Brownlees, C., Engle, R., Bryan, K., 2012. A Practical Guide to Volatility Forecasting Through Calm and Storm. *Journal of Risk*, 14, 3-22.

Christiansen, C., Schmeling, M., Schrimpf, A., 2012. A comprehensive look at financial volatility prediction by economic variables. *Journal of Applied Econometrics*, 27, 956-977.

Conrad, C., Loch, K., 2015. Anticipating Long-Term Stock Market Volatility. *Journal of Applied Econometrics*, 30, 1090-1114.

Conrad, C., Loch, K., Rittler, D., 2014. On the Macroeconomic Determinants of Long-Term Volatilities and Correlations in U.S. Stock and Crude Oil Markets. *Journal of Empirical Finance*, 29, 26-40.

Conrad, C., Schienle, M., 2018. Testing for an Omitted Multiplicative Long-Term Component in GARCH Models. Available at SSRN: https://ssrn.com/abstract=2631976.

Corsi, F., M., 2009. A Simple Approximate Long-Memory Model of Realized Volatility. Department of Economics, *Journal of Financial Econometrics*, 7, 174-196.

Diebold, F. X., Mariano, R. S., 1995. Comparing Predictive Accuracy. *Journal of Business and Economic Statistics*, 13, 253-263.

Ding, Z., Granger, C., 1996. Modeling Volatility Persistence of Speculative Returns: A New Approach. *Journal of Econometrics*, 73, 185-215.

Dominicy, Y., Vander Elst, H., 2015. Macro-Driven VAR Forecasts: From Very High to Very Low-Frequency Data. Available at SSRN: https://ssrn.com/abstract=2701256.

Dorion, C., 2016. Option Valuation with Macro-Finance Variables. *Journal of Financial and Quantitative Analysis*, 51, 1359-1389.

Engle, R. F., Ghysels, E., Sohn, B., 2013. Stock Market Volatility and Macroeconomic Fundamentals. *Review of Economics and Statistics*, 95, 776-797.

Engle, R., F., Lee, G., 1999. A Long-Run and Short-Run Component Model of Stock Return Volatility. In R. Engle and H. White (eds.), *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive WJ Granger*, pp. 475-497. Oxford University Press.

Engle, R., F., Rangel, J., 2008. The Spline-GARCH Model for Low-Frequency Volatility and its Global Macroeconomic Causes. *Review of Financial Studies*, 21, 1187-1222.

Francq, C., Thieu, L. Q., 2015. QML Inference for Volatility Models with Covariates. MPRA Paper 63198, University Library of Munich, Germany.

Glosten, L.R., Jagannathan, R., Runkle, D.E., 1993. On the Relation between the Expected Value and the Volatility of Nominal Excess Return on Stocks. *Journal of Finance*, 48, 1779-1801.

Han, H., 2015. Asymptotic Properties of GARCH-X Processes. *Journal of Financial Econometrics*, 13, 188-221.

Han, H., Kristensen, D., 2015. Semiparametric Multiplicative GARCH-X Model: Adopting Economic Variables to Explain Volatility. Working Paper.

Han, H., Park, J. Y., 2014. GARCH with Omitted Persistent Covariate. *Economics Letters*, 124, 248-254.

Hansen P. R., Lunde A., 2006. Consistent Ranking of Volatility Models. *Journal of Econometrics*, 131, 97-121.

Hansen P. R., Lunde A., Nason J. M., 2011. The Model Confidence Set. *Econometrica*, 79, 453-497.

Heber, G., Lunde, A., Shephard, N., Sheppard, K. K., 2009. Oxford-Man Institute's Realized Library. Oxford-Man Institute. University of Oxford.

Hillebrand, E., 2005. Neglecting Parameter Changes in GARCH Models. *Journal of Econometrics*, 129, 121-138.

Karanasos, M., 1999. The Second Moment and the Autocovariance Function of the Squared Errors of the GARCH Model. *Journal of Econometrics*, 90, 63-76.

Lindblad, A., 2017. Sentiment Indicators and Macroeconomic Data as Drivers for Low-Frequency Stock Market Volatility. Helsinki Center of Economic Research, Discussion Paper No. 413.

Kleen, O., 2017. alfred: Downloading Time Series from ALFRED Database for Various Vintages. R package, available on CRAN: https://cran.r-project.org/package=alfred.

Kleen, O., 2018. mfGARCH: Mixed-Frequency GARCH Models. R package, available on CRAN: https://cran.r-project.org/package=mfGARCH.

Opschoor, A., van Dijk, D., van der Wel, M., 2014. Predicting Volatility and Correlations with Financial Conditions Indexes. *Journal of Empirical Finance*, 29, 435-447.

Pan, Z., Wang, Y., Wu, C., Yin, L., 2017. Oil Price Volatility and Macroeconomic Fundamentals: A Regime Switching GARCH-MIDAS model. *Journal of Empirical Finance*, 43, 130-142.

Patton, A., 2011. Volatility Forecast Comparison Using Imperfect Volatility Proxies. *Journal of Econometrics*, 160, 246-256.

Patton, A., 2016. Comparing Possibly Misspecified Forecasts. Working Paper.

Patton, A.J., Sheppard, K., 2009. Evaluating Volatility and Correlation Forecasts. In T. Mikosch, J.-P. Kreiß, R.A. Davis, T.G. Andersen (eds.), *Handbook of Financial Time Series*, pp. 801-838. Springer Berlin Heidelberg.

Paye, B. S., 2012. 'Déja Vol': Predictive regressions for aggregate stock market volatility using macroeconomic variables. *Journal of Financial Economics*, 106, 527-546.

Wang, F., Ghysels, E., 2015. Econometric Analysis of Volatility Component Models. *Econometric Theory*, 31, 362-393.

A Proofs

Proof of Proposition 1. The proof follows directly by applying the mutual independence of $g_{i,t}$, τ_t and $Z_{i,t}$ and by noting that Assumption 3 implies that $\mathbf{E}[\tau_t^2]/\mathbf{E}[\tau_t]^2 > 1$ if τ_t is non-constant. □

Proof of Proposition 2. First, note that under Assumptions 1, 2 and 3 the covariance $\mathbf{Cov}(r_s^2, r_{s-k}^2)$ exists for every $k \in \mathbb{N}$ and is time-invariant. In the proof, we use that τ_s and g_s are independent covariance stationary processes and that Z_s are *i.i.d.* innovations.

$$
\begin{aligned}
\rho_k^{MG} &= \frac{\mathbf{Cov}(r_s^2, r_{s-k}^2)}{\sqrt{\mathbf{Var}(r_s^2)}\sqrt{\mathbf{Var}(r_{s-k}^2)}} \\
&= \frac{\mathbf{E}[\tau_s\tau_{s-k}]\mathbf{E}[g_s Z_s^2 g_{s-k} Z_{s-k}^2] - \mathbf{E}[\tau_s]\mathbf{E}[\tau_{s-k}]}{\mathbf{Var}(r_s^2)} \\
&= \frac{\mathbf{E}[\tau_s\tau_{s-k}]\mathbf{E}[g_s Z_s^2 g_{s-k} Z_{s-k}^2] - \mathbf{E}[\tau_s\tau_{s-k}] + \mathbf{E}[\tau_s\tau_{s-k}] - \mathbf{E}[\tau_s]\mathbf{E}[\tau_{s-k}]}{\mathbf{Var}(r_s^2)} \\
&= \frac{\mathbf{E}[\tau_s\tau_{s-k}] - \mathbf{E}[\tau_s]^2}{\mathbf{Var}(r_s^2)} + \frac{\left(\mathbf{E}[g_s Z_s^2 g_{s-k} Z_{s-k}^2] - \mathbf{E}[g_s]\mathbf{E}[g_{s-k}]\right)\mathbf{E}[\tau_s\tau_{s-k}]}{\mathbf{Var}(r_s^2)} \\
&= \frac{\mathbf{Cov}(\tau_s, \tau_{s-k})}{\mathbf{Var}(r_s^2)} + \frac{\mathbf{Cov}(g_s Z_s^2, g_{s-k} Z_{s-k}^2)(\mathbf{Cov}(\tau_s, \tau_{s-k}) + \mathbf{E}[\tau_s^2])}{\mathbf{Var}(r_s^2)} \\
&= \rho_k^{\tau}\frac{\mathbf{Var}(\tau_s)}{\mathbf{Var}(r_s^2)} + \rho_k^{GA}\frac{\left(\rho_k^{\tau}\mathbf{Var}(\tau_s) + \mathbf{E}[\tau_s]^2\right)\mathbf{Var}(g_t Z_t^2)}{\mathbf{Var}(r_t^2)}
\end{aligned}
$$

□

Proof of Proposition 3. The expression for R_k^2 and its limit in $\mathbf{E}[\tau_{t+1}^2]$ are derived by using the independence of $g_{k,t+1|t}$ and τ_{t+1}. As $k \to \infty$, $\mathbf{Var}(g_{k,t+1|t}^2)$ decreases monotonically to zero implying that $\mathbf{E}[g_{k,t+1|t}^2]$ decreases monotonically towards one because $\mathbf{E}[g_{k,t+1|t}] = 1$. This implies eq. (20). The limit with respect to $\mathbf{E}[\tau_{t+1}^2]$ follows from rewriting

$$
R_k^2 = \frac{\mathbf{Var}(h_{k,t+1|t})}{\mathbf{Var}(r_{k,t+1}^2)} = \frac{\mathbf{E}[g_{k,t+1|t}^2] - \frac{\mathbf{E}[\tau_{t+1}]^2}{\mathbf{E}[\tau_{t+1}^2]}}{\mathbf{E}[g_{k,t+1}^2]\kappa - \frac{\mathbf{E}[\tau_{t+1}]^2}{\mathbf{E}[\tau_{t+1}^2]}}.
$$

Hence, $R_k^2 \to \frac{\mathbf{E}[g_{k,t+1|t}^2]}{\mathbf{E}[g_{k,t+1}^2]\kappa}$ as $\mathbf{E}[\tau_{t+1}^2] \to \infty$.

□

Proof of Lemma 1. Using eq. (4), we obtain

$$
R_1^2 = \frac{\mathbf{Var}(g_t\tau_t)}{\mathbf{Var}(r_t^2)} = \frac{\mathbf{E}[g_t^2]\mathbf{E}[\tau_t^2] - \mathbf{E}[\tau_t]^2}{\mathbf{E}[g_t^2]\mathbf{E}[\tau_t^2]\kappa - \mathbf{E}[\tau_t]^2}
$$

$$= \frac{(1-(\alpha+\gamma/2+\beta)^2)\mathbf{E}[\tau_t^2] - (1-(\alpha+\gamma/2)^2\kappa - 2(\alpha+\gamma/2)\beta - \beta^2)\mathbf{E}[\tau_t]^2}{(1-(\alpha+\gamma/2+\beta)^2)\mathbf{E}[\tau_t^2]\kappa - (1-(\alpha+\gamma/2)^2\kappa - 2(\alpha+\gamma/2)\beta - \beta^2)\mathbf{E}[\tau_t]^2}.$$

The limit of R_1^2 follows from the proof of Proposition 3, when noting that $\mathbf{E}[g_{k,t+1|t}^2] = \mathbf{E}[g_{k,t+1}^2]$.

□

B Tables

Table 1: Monte-Carlo parameter estimates.

		α	β	m	θ	w_2	K	$\kappa-3$	VR(X)
Monthly τ_t	M-GARCH	0.060	0.907	0.092	0.326	4.791	36	-0.009	21.96
		[0.055,0.065]	[0.897,0.915]	[0.043,0.144]	[0.246,0.402]	[2.629,9.624]		[-0.059,0.040]	
		0.060	0.907	0.091	0.316	3.194	24	-0.009	21.84
		[0.054,0.065]	[0.897,0.915]	[0.044,0.143]	[0.236,0.390]	[1.708,6.677]		[-0.059,0.040]	
		0.060	0.907	0.091	0.292	1.571	12	-0.008	21.08
		[0.055,0.065]	[0.898,0.916]	[0.044,0.142]	[0.218,0.363]	[1.000,3.729]		[-0.059,0.041]	
	GARCH	0.060	0.914	0.110	—	—	—	0.002	—
		[0.055,0.066]	[0.905,0.922]	[0.043,0.182]				[-0.047,0.053]	
Daily τ_t (low VR)	M-GARCH	0.060	0.908	-0.105	0.306	4.965	264	-0.012	34.16
		[0.055,0.065]	[0.899,0.916]	[-0.144,-0.059]	[0.258,0.360]	[3.685,6.817]		[-0.055,0.035]	
		0.060	0.908	-0.104	0.290	2.366	132	-0.012	33.85
		[0.055,0.065]	[0.899,0.917]	[-0.144,-0.059]	[0.245,0.337]	[1.744,3.311]		[-0.055,0.035]	
		0.060	0.909	-0.100	0.247	1.237	66	-0.010	30.08
		[0.055,0.065]	[0.900,0.917]	[-0.140,-0.058]	[0.210,0.284]	[1.000,1.830]		[-0.053,0.037]	
	GARCH	0.063	0.914	-0.065	—	—	—	0.013	—
		[0.058,0.068]	[0.906,0.921]	[-0.118,-0.014]				[-0.035,0.061]	
Daily τ_t (high VR)	M-GARCH	0.060	0.908	-0.105	0.301	4.980	264	-0.011	74.93
		[0.054,0.065]	[0.898,0.917]	[-0.147,-0.059]	[0.281,0.323]	[4.430,5.584]		[-0.057,0.035]	
		0.060	0.908	-0.104	0.286	2.383	132	-0.010	74.12
		[0.054,0.066]	[0.899,0.917]	[-0.146,-0.059]	[0.268,0.306]	[2.097,2.688]		[-0.056,0.037]	
		0.062	0.908	-0.095	0.246	1.231	66	-0.002	67.38
		[0.056,0.067]	[0.899,0.917]	[-0.139,-0.051]	[0.229,0.261]	[1.067,1.435]		[-0.047,0.047]	
	GARCH	0.076	0.915	0.135	—	—	—	0.086	—
		[0.072,0.081]	[0.909,0.920]	[0.024,0.258]				[0.036,0.143]	

Notes: The table reports the median parameter estimates and the corresponding inter-quartile ranges across the 2000 Monte-Carlo simulations. We provide results for both daily- and monthly long-term components. The parameter estimates are based on (the first) 20 years of observations (i.e. the in-sample period). In both long-term components, see eq. (5), we choose $\theta = 0.3$, $w_1 = 1$ and $w_2 = 5$. We use $m = 0.1$ in the monthly τ_t and $m = -0.1$ in the daily τ_t. The long-term component is assumed to depend on the recent 36 monthly observations or 264 daily ones. The covariate X_t is modeled as an AR(1)-process, i.e. $X_s = \psi X_{s-1} + \xi_s, \xi_s \overset{i.i.d.}{\sim} \mathcal{N}(0,\ \sigma_\xi^2)$, in which we choose $\psi = 0.95$, $\sigma_\xi^2 = 0.2^2$ for a monthly-, and $\psi = 0.98$, $\sigma_\xi^2 = 0.2^2$ (low VR) or $\sigma_\xi^2 = 0.5^2$ (high VR) for a daily-varying τ_t. The parameters of the short-term component are in both cases given by $\alpha = 0.06$, $\beta = 0.91$, $\gamma = 0$ and $Z_{i,t} \sim \mathcal{N}(0,1)$. K denotes the lag length imposed when estimating the different specifications. The column '$\kappa-3$' presents the average excess kurtosis of the standardized residuals from each model. VR(X) = $\mathbf{Var}(\log(\tau_s))/\mathbf{Var}(\log(\sigma_s^2))$.

Table 2: DM tests for daily τ_t (high VR).

	SE			
	$k=1$	$k=5$	$k=10$	$k=20$
Full sample	0.814	0.771	0.650	0.514
Low regime sample	0.762	0.692	0.565	0.434
Normal regime	0.930	0.890	0.798	0.664
High regime	0.614	0.578	0.427	0.291
	QLIKE			
Full sample	0.983	0.970	0.923	0.840
Low regime sample	0.789	0.727	0.608	0.443
Normal regime	0.964	0.929	0.850	0.725
High regime	0.816	0.761	0.613	0.442

Notes: The numbers are empirical rejection frequencies of the Diebold-Mariano test for equal predictive accuracy in favor of the M-GARCH model at the nominal 5% level. k denotes the forecast horizon. Diebold-Mariano tests in the low- and high-volatility regime are based on 60 non-overlapping observations and on 120 observations in the normal regime. The tests are based on 2000 Monte-Carlo replications.

Table 3: Summary statistics.

Variable	Freq.	Start	Obs.	Min.	Max.	Mean	Median	Sd.	Skew.	Kurt.	AC(1)
Stock market data											
S&P 500 returns	d	1971	11644	-22.90	10.96	0.03	0.04	1.07	-1.03	28.72	0.00
RV	d	2000	4287	0.02	77.48	1.19	0.55	2.59	11.14	225.61	0.67
Explanatory variables											
$\sqrt{\widetilde{\mathrm{RV}}(22)}$	d	1988	7349	0.29	5.54	0.97	0.82	0.56	3.00	17.60	0.99
$\mathrm{VIX}/\sqrt{252}$	d	1990	6841	0.59	5.09	1.24	1.12	0.49	2.11	10.74	0.98
NFCI	w	1973	2305	-1.03	4.71	0.00	-0.37	1.00	1.92	6.48	0.99
NAI	m	1971	554	-5.09	2.72	-0.00	0.07	1.00	-1.22	6.88	0.67
Δ IP	m	1971	554	-4.40	2.38	0.18	0.22	0.72	-1.24	8.85	0.36
Δ Housing	m	1971	554	-30.67	25.67	-0.07	-0.13	8.05	-0.03	3.78	-0.32

Notes: The table presents summary statistics for the different variables, whereby the column Start indicates the year of the first observation for each variable. The data ends in 2017:M2. The column frequency (Freq.) indicates whether the data is observed on a daily (d), weekly (w) or monthly (m) basis. The reported statistics include the number of observations (Obs.), the minimum (Min.) and maximum (Max.), the mean and median, standard deviation (Sd.), Skewness (Skew.), Kurtosis (Kurt.) and the first-order autocorrelation coefficient (AC(1)). We define $\widetilde{RV}_{i,t}(22) = 1/22 \sum_{j=0}^{21} r^2_{i-j,t}$. Changes in industrial production and housing starts are measured in month-over-month log differences, i.e. $\Delta X_t = 100 \cdot (\ln(X_t) - \ln(X_{t-1}))$.

Table 4: LM-Test for misspecification of GJR-GARCH(1,1).

X_t	VIX/$\sqrt{252}$	$\sqrt{\widetilde{\text{RV}}(22)}$	NFCI	NAI	Δ IP	Δ Hous
$K=1$	70.62 [<0.01]	2.44 [0.12]	17.64 [<0.01]	18.52 [<0.01]	10.18 [<0.01]	0.43 [0.51]
$K=2$	78.40 [<0.01]	6.06 [0.05]	19.05 [<0.01]	20.80 [<0.01]	12.98 [<0.01]	0.64 [0.73]

Notes: The table reports the test statistics and the corresponding p-values of the Conrad and Schienle (2018) misspecification test for one-component GJR-GARCH(1,1) models. The test is implemented using either one ($K=1$) or two ($K=2$) lags of the explanatory variable X_t. For VIX/$\sqrt{252}$ and $\sqrt{\widetilde{\text{RV}}(22)}$ the test is based on daily data from 1990 onwards, for NFCI, NAI, Δ IP, and Δ Housing starts the test is based on weekly/monthly data from 1973 onwards.

Table 5: Full sample estimation results for M-GARCH models.

	μ	α	β	γ	m	θ	w_1	w_2	K	LLF	BIC	VR(X)
Daily τ_t												
$\sqrt{\widetilde{\text{RV}}(22)}$	0.028*** (0.009)	0.000 (0.010)	0.843*** (0.020)	0.184*** (0.024)	−1.202*** (0.115)	1.124*** (0.084)	1	3.240*** (0.767)	264	−9288	18638	42.88
VIX/$\sqrt{252}$	0.018* (0.009)	0.000 (0.012)	0.858*** (0.028)	0.091*** (0.018)	−2.034*** (0.097)	0.430*** (0.021)	1	3.443*** (0.743)	3	−8883	17827	75.31
Weekly τ_t												
NFCI	0.027*** (0.008)	0.013** (0.005)	0.898*** (0.017)	0.122*** (0.024)	−0.062 (0.104)	0.301*** (0.041)	1	2.314*** (0.720)	52	−14252	28570	16.39
Monthly τ_t												
NAI	0.028*** (0.008)	0.017*** (0.005)	0.901*** (0.016)	0.115*** (0.022)	−0.054 (0.118)	−0.360*** (0.069)	1	8.580** (3.744)	36	−14299	28663	14.75
Δ IP	0.028*** (0.008)	0.017*** (0.005)	0.904*** (0.015)	0.114*** (0.022)	0.082 (0.123)	−0.703*** (0.154)	1	4.689*** (1.090)	36	−14302	28669	12.19
Δ Housing	0.028*** (0.008)	0.018*** (0.005)	0.898*** (0.017)	0.117*** (0.023)	−0.072 (0.110)	−0.229*** (0.042)	1.645* (0.858)	2.616** (1.251)	36	−14290	28655	19.66
GARCH	0.029*** (0.007)	0.019*** (0.005)	0.912*** (0.014)	0.103*** (0.020)	−0.066 (0.133)	—	—	—	—	−15088	30222	—

Notes: Estimation results for M-GARCH models are reported for four different covariates varying at different frequencies. The estimates are based on daily return data stretching from 1971:M1 to 2017:M2. Estimation using NFCI, NAI, IP and housing starts begins in 1974:M1 and when using realized variance or VIX in 1990:M1. For all covariates except housing starts a restricted weighting scheme is chosen ($w_1=1$). Bollerslev-Wooldridge robust standard errors are reported in parenthesis where significance at the 1, 5, 10 % level is indicated by ***, ** and *. LLF is the value of the maximized log-likelihood function and BIC is the Bayesian Information Criterion. The variance ratio $\text{VR}(X) = \mathbf{Var}(\log(\tau_M^X))/\mathbf{Var}(\log(\sigma_M^X))$ is calculated on monthly aggregates.

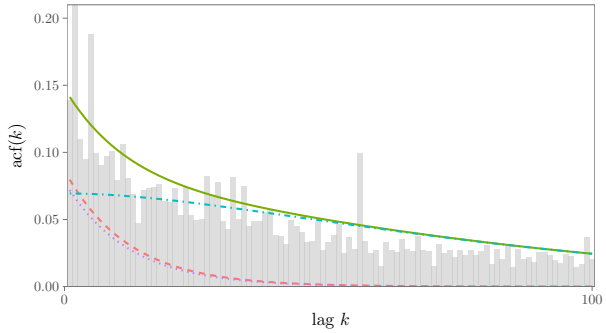
Table 6: Forecast error statistics.

Days	Model	$\left\lvert\hat{\sigma}^2_{(\cdot)} - \hat{h}_{(\cdot)}\right\rvert$	$\frac{\hat{\sigma}^2_{(\cdot)}}{\hat{h}_{(\cdot)}}$	Month	Model	$\left\lvert\hat{\sigma}^2_{(\cdot)} - \hat{h}_{(\cdot)}\right\rvert$	$\frac{\hat{\sigma}^2_{(\cdot)}}{\hat{h}_{(\cdot)}}$
1	$\sqrt{\widetilde{\text{RV}}(22)}$	0.63	0.65	1	$\sqrt{\widetilde{\text{RV}}(22)}$	13.03	0.69
	VIX/$\sqrt{252}$	0.54	0.78		VIX/$\sqrt{252}$	10.33	0.97
	NFCI	0.55	0.79		NFCI	10.19	0.89
	NAI	0.58	0.75		NAI	11.66	0.82
	Δ IP	0.59	0.76		Δ IP	11.17	0.83
	Δ Housing	0.57	0.78		Δ Housing	10.71	0.87
	GARCH	0.58	0.74		GARCH	11.19	0.82
	HAR	0.41	1.22		HAR	7.82	1.20
	No-change	0.50	1.35		No-change	11.99	1.87
8	$\sqrt{\widetilde{\text{RV}}(22)}$	4.54	0.69	2	$\sqrt{\widetilde{\text{RV}}(22)}$	16.04	0.67
	VIX/$\sqrt{252}$	3.79	0.89		VIX/$\sqrt{252}$	12.18	1.06
	NFCI	3.68	0.85		NFCI	11.95	0.89
	NAI	4.10	0.80		NAI	13.91	0.79
	Δ IP	3.92	0.81		Δ IP	13.65	0.81
	Δ Housing	3.80	0.84		Δ Housing	12.49	0.86
	GARCH	3.93	0.80		GARCH	13.94	0.79
	HAR	2.79	1.19		HAR	10.78	1.14
	No-change	4.05	1.63		No-change	14.86	2.27
15	$\sqrt{\widetilde{\text{RV}}(22)}$	8.69	0.69	3	$\sqrt{\widetilde{\text{RV}}(22)}$	17.27	0.67
	VIX/$\sqrt{252}$	7.07	0.94		VIX/$\sqrt{252}$	13.36	1.10
	NFCI	6.92	0.88		NFCI	12.47	0.89
	NAI	7.83	0.82		NAI	14.66	0.77
	Δ IP	7.46	0.82		Δ IP	14.49	0.81
	Δ Housing	7.23	0.86		Δ Housing	12.67	0.84
	GARCH	7.48	0.82		GARCH	15.07	0.77
	HAR	5.25	1.20		HAR	12.08	1.08
	No-change	7.99	1.78		No-change	16.24	2.28

Notes: In this table the average absolute forecast errors and the average standardized forecast errors are reported. The left side displays figures for cumulative 1-, 5- and 15-days ahead forecasts. The right side presents forecast errors for 1-, 2- and 3-month ahead. Note that the one-month ahead forecast corresponds to a cumulative 22-days ahead forecast. The out-of-sample period spreads 2010:M1 - 2017:M2.

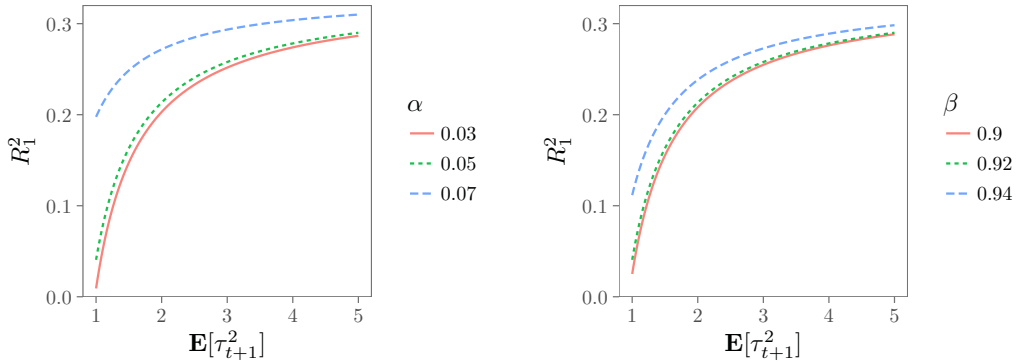
C Figures

Figure 1: Autocorrelation function.



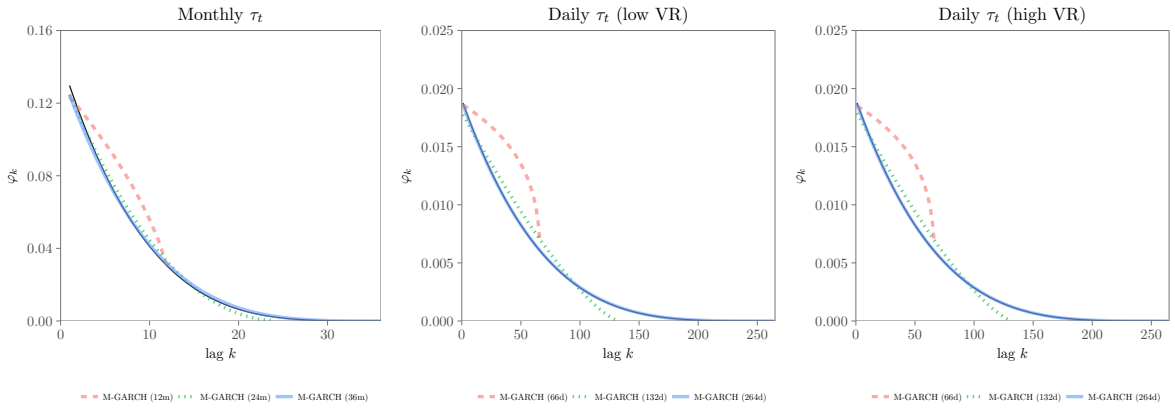
Notes: We depict the ACF of a multiplicative GARCH model (green, solid), the first (blue, dot-dashed) and second term (purple, dotted) in eq. (9), as well as the ACF of the nested GARCH(1,1) model (red, dashed). The long-term component is defined as in eq. (5) with $m = -0.1$, $\theta = 0.3$, $w_1 = 1, w_2 = 6$ and $K = 264$ smoothing an exogenous covariate $X_s = \psi X_{s-1} + \xi_s$, $\xi_s \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\xi^2)$, where $\psi = 0.98$ and $\sigma_\xi^2 = 0.35^2$. The GARCH(1,1) parameters are $\alpha = 0.06$, $\beta = 0.9$ and $\gamma = 0$. Moreover, we set $\kappa = 3$. Bars in light gray display the empirical autocorrelation of squared S&P 500 returns in between 1971:M1 and 2017:M2, for details see Section 4.

Figure 2: R_1^2 as a function of $\mathbf{E}[\tau_{t+1}^2]$.



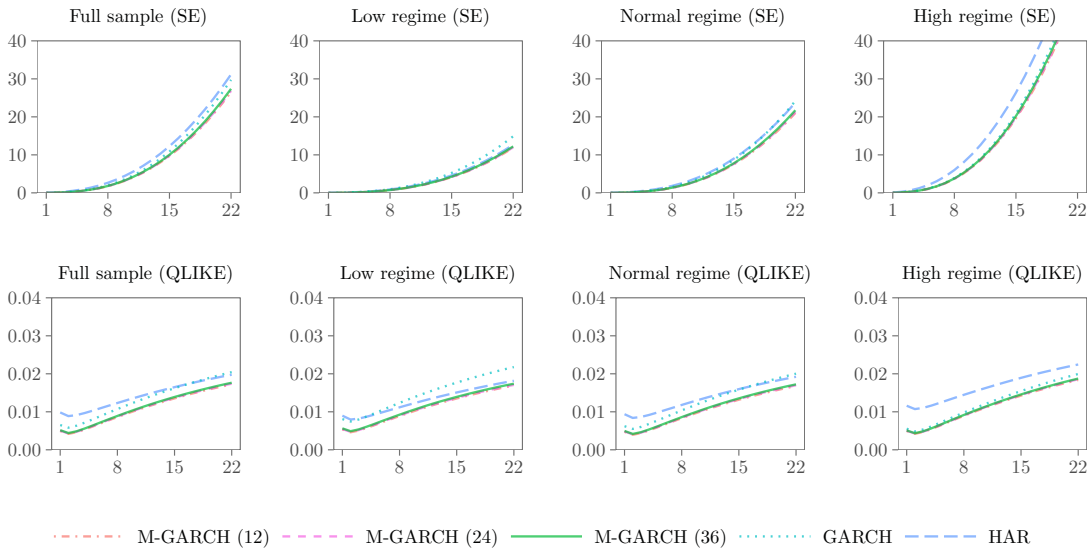
Notes: We depict the population R_1^2 of a Mincer-Zarnowitz regression as a function of $\mathbf{E}[\tau_{t+1}^2]$. In the left plot, β equals 0.92. In the right plot, we choose $\alpha = 0.05$. In all cases, we set $\mathbf{E}[\tau_{t+1}] = 1$, $\gamma = 0$ and $\kappa = 3$.

Figure 3: Weighting schemes implied by median parameter estimates.



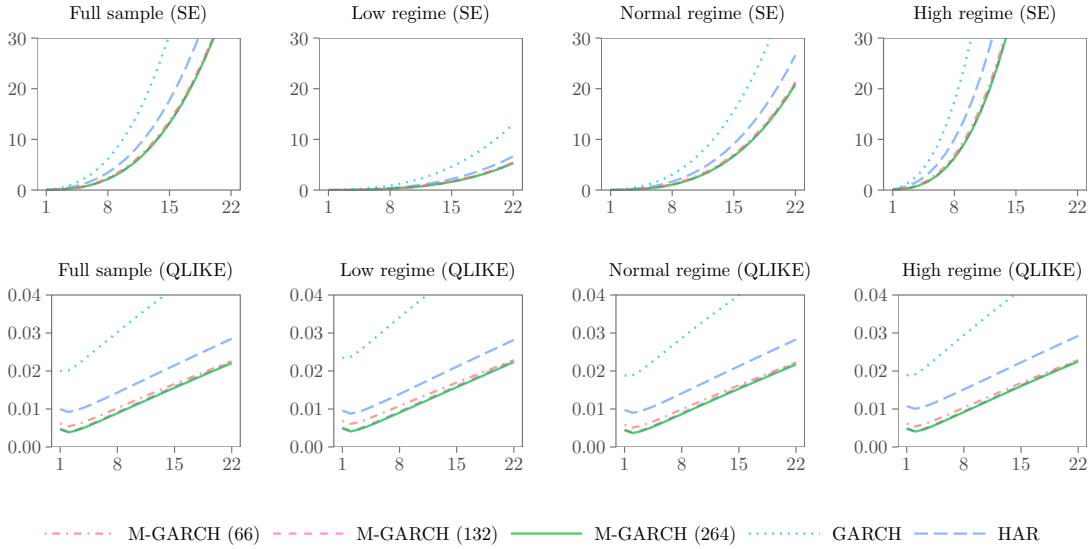
Notes: Estimated Beta weighting schemes (see eq. (6)) as implied by the median parameter estimates reported in Table 1. The blue (solid) line corresponds to the case of a correctly specified model whereas the green (dotted) and red (dashed) line correspond to models with K being too small. The true weighting scheme is depicted in dark gray (barely visible behind blue line).

Figure 4: Monthly τ_t - SE and QLIKE loss across regimes.



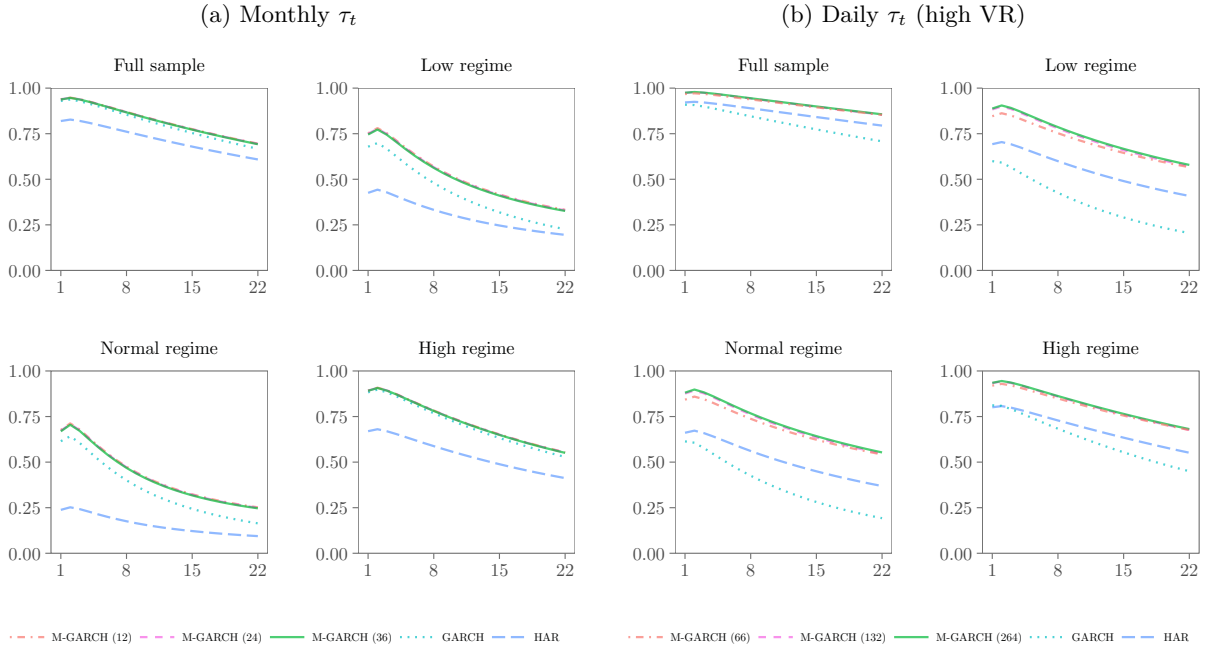
Notes: Average SE and QLIKE loss evaluated at different regimes in the case of a monthly-varying τ_t. In each simulation, we calculate out-of-sample SE and QLIKE loss as well as losses depending on three volatility regimes for 1–22 days ahead; forecasts being issued at a day for which the daily realized volatility is below the empirical 25% quantile (low regime), between the 25% and 75% quantile (normal regime) or above the 75% quantile (high regime). Averages are taken across 2000 simulations.

Figure 5: Daily τ_t (high VR) - SE and QLIKE loss across regimes.



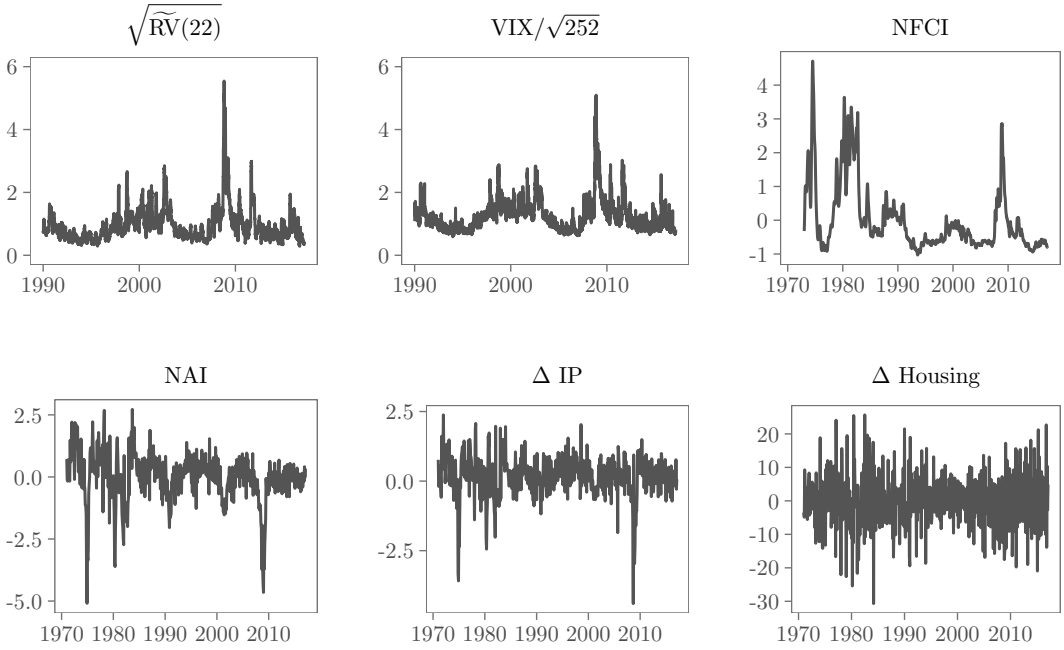
Notes: Average SE and QLIKE loss evaluated at different regimes in the case of a daily-varying τ_t with a high variance ratio. In each simulation, we calculate out-of-sample SE and QLIKE loss as well as losses depending on three volatility regimes for 1–22 days ahead; forecasts being issued at a day for which the daily realized volatility is below the empirical 25% quantile (low regime), between the 25% and 75% quantile (normal regime) or above the 75% quantile (high regime). Averages are taken across 2000 simulations.

Figure 6: Mincer-Zarnowitz $R^2_{1:k}$.



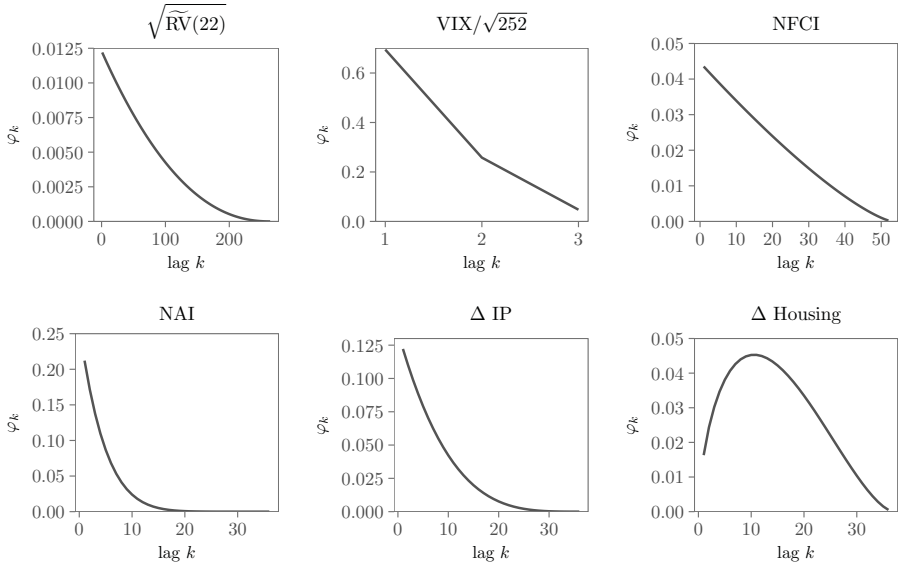
Notes: Average coefficients of determination of Minzer-Zarnowitz regressions for different volatility regimes. Hereby, we regress the cumulative realized variance on the cumulative forecast. The upper panel displays the case of the monthly-varying long-term component and the lower panel the case of a daily-varying long-term component with a high variance ratio. In each simulation, we calculate out-of-sample $R^2_{1:k}$s depending on three volatility regimes; forecasts being issued at a day for which the daily realized volatility is below the empirical 25% quantile (low regime), between the 25% and 75% quantile (normal regime) or above the 75% quantile (high regime). Averages are taken across 2000 simulations.

Figure 7: Time series of explanatory variables.



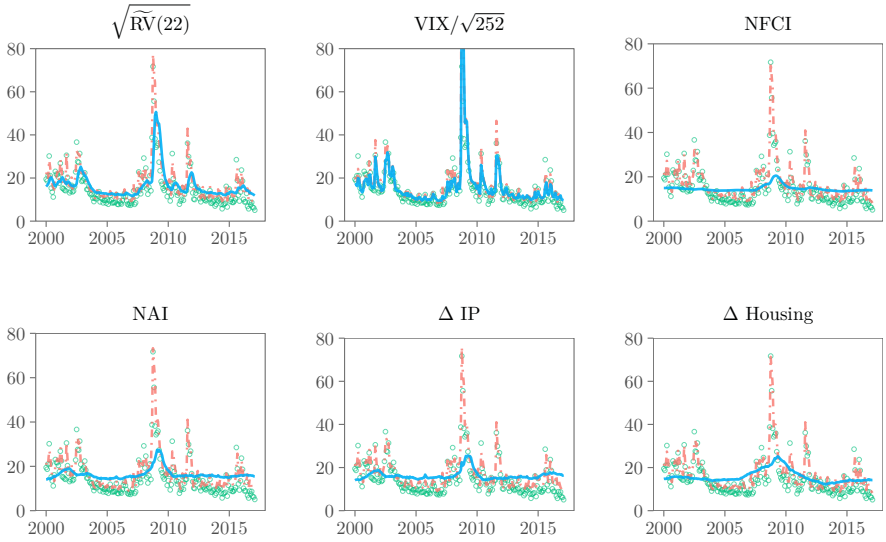
Notes: Daily financial data for the 1990:M1 to 2017:M2 period and macroeconomic data for the 1971:M1 to 2017:M2 period. See Table 3 for definition and descriptive statistics of those variables.

Figure 8: Weighting schemes for different explanatory variables.



Notes: For each explanatory variable, the estimated Beta weighting schemes (see eq. (6)) based on full-sample estimates are depicted. For all variables except housing starts, we impose the restriction $w_1 = 1$. The corresponding parameters are reported in Table 5.

Figure 9: Monthly conditional volatility components.



Notes: The figure shows the monthly long-run volatility components $\sqrt{\tau_M}$ (blue, solid) and the monthly conditional volatilities $\sqrt{g_M \tau_M}$ (red, dash-dotted) for all GARCH-MIDAS models in between 2000:M1 and 2017:M2. Circles correspond to realized volatilities. Volatility is measured on an annualized scale.

Figure 10: Real-time forecasting with macroeconomic releases.

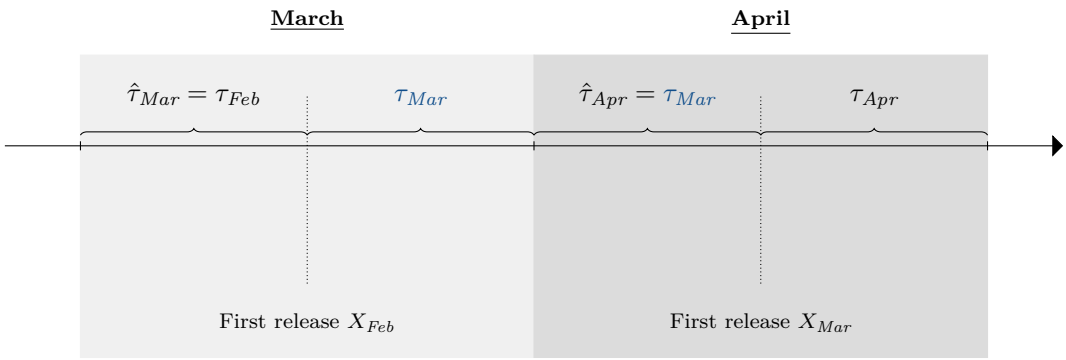
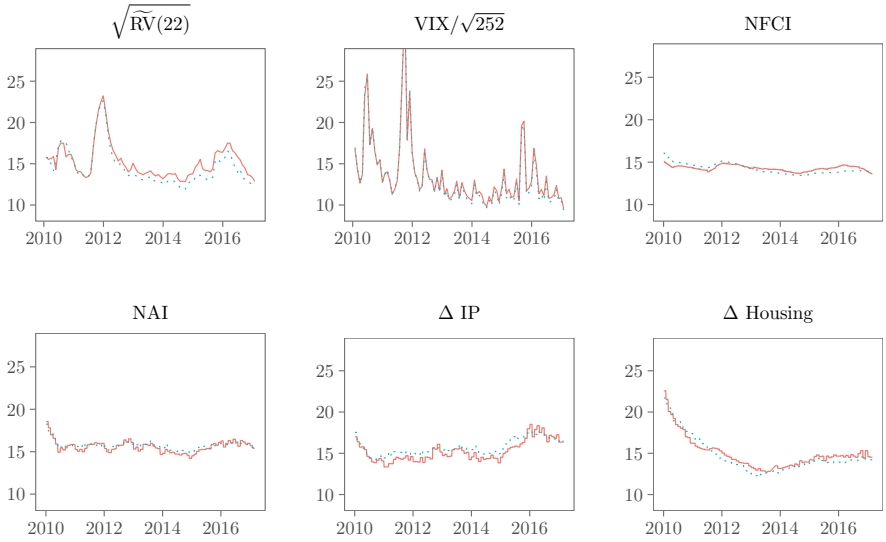
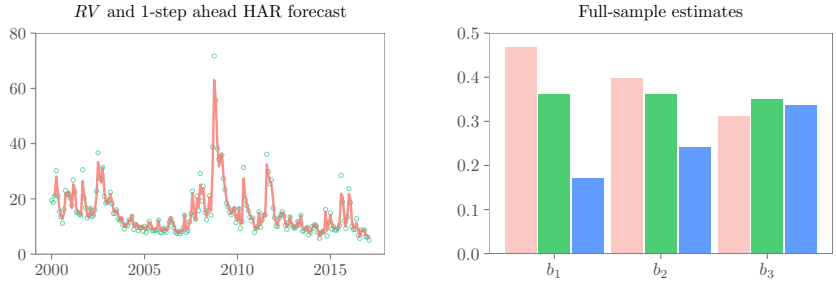


Figure 11: Comparison of rolling window and full-sample long-term components.



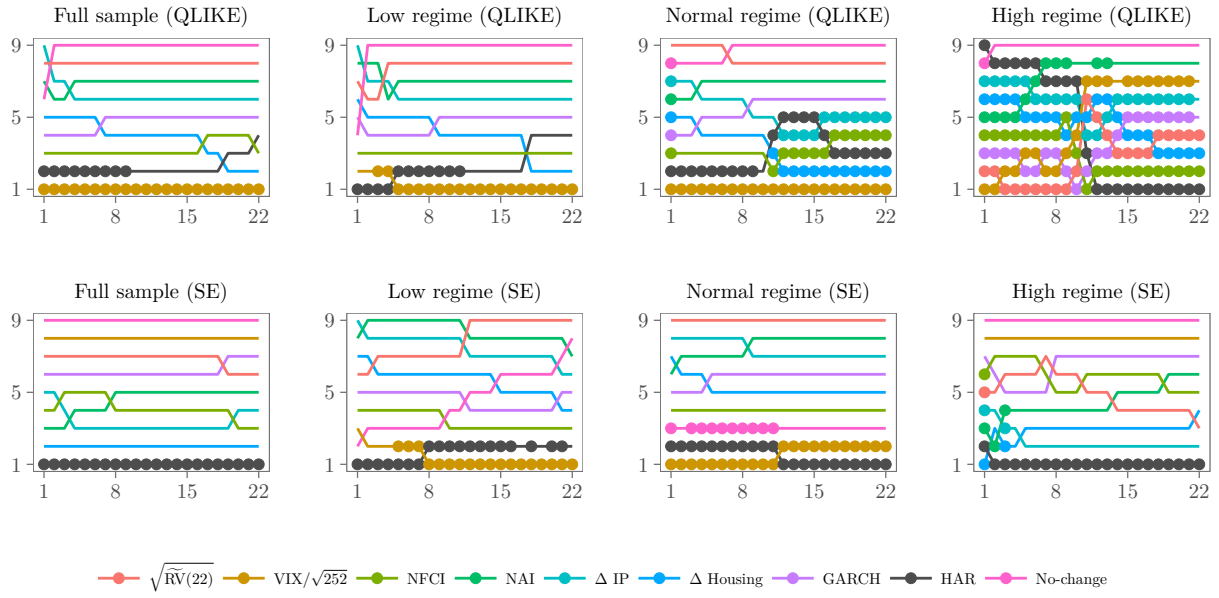
Notes: For each explanatory variable, the long-term volatility components, $\sqrt{\tau_t}$, are depicted for the period 2010:M1 to 2017:M2. The long-term component obtained from the full-sample estimates is given in green (dotted) along the real-time estimates of $\sqrt{\tau_t}$ in red (solid). Volatilities are presented on an annualized scale.

Figure 12: HAR model.



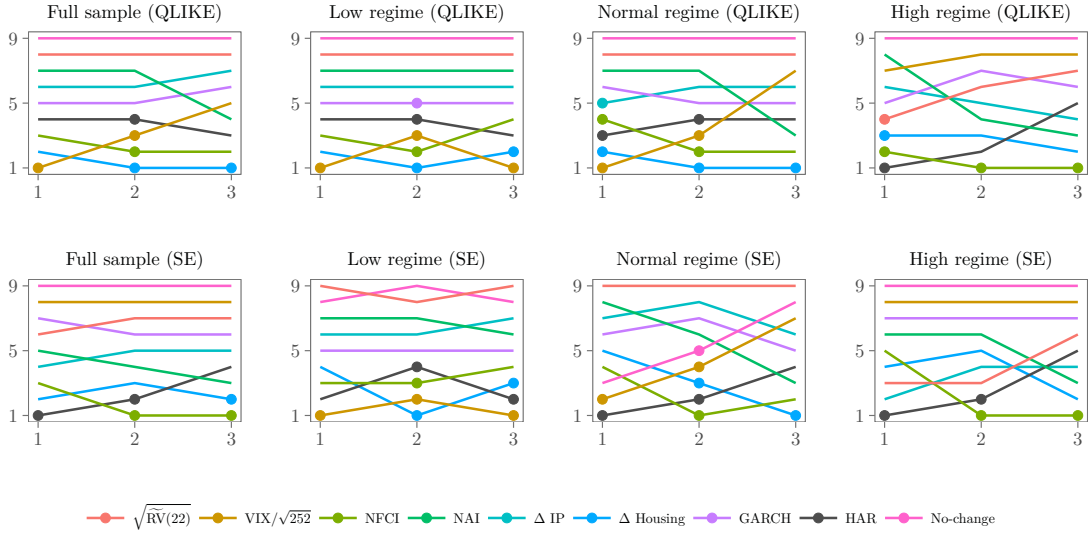
Notes: On the left side, monthly aggregates of one-day ahead volatility predictions from the HAR model (red, solid) along with monthly realized volatilities (green, circles) are depicted. Both volatilities are presented on an annualized scale. On the right side, the parameter estimates for the HAR model, see eq. (24), are reported based on the full-sample. Red bars correspond to a HAR forecasting one day ahead, the green bars to forecasting cumulative realized variances five days ahead and the blue bars to predicting cumulative 22-days ahead. Values for *RV* are used in between 2000:M1 – 2017:M2.

Figure 13: MCS ranking for SE and QLIKE loss. Cumulative 1- to 22-days ahead forecasts.



Notes: Model Confidence Set rankings for conditional variance forecasts aiming at periods of 1- to 22-days ahead are depicted. Models that are included in the MCS are indicated by a dot. The ranking is based on the average loss in the particular regime. We consider two loss functions, SE and QLIKE loss and three different volatility regimes; forecasts being issued at a day for which the daily realized volatility is below the empirical 25% quantile (low regime), between the 25% and 75% quantile (normal regime) or above the 75% quantile (high regime). The out-of-sample evaluation period spreads 2010:M1 to 2017:M2.

Figure 14: MCS ranking for SE and QLIKE loss, 1-, 2-, 3-months ahead forecasts.



Notes: Model Confidence Set rankings for 1, 2 and 3 month ahead conditional variance forecasts are depicted. Models that are included in the MCS are indicated by a dot. The ranking is based on the average loss in the particular regime. We consider two loss functions, SE and QLIKE loss and three different volatility regimes; forecasts being issued at a day for which the daily realized volatility is below the empirical 25% quantile (low regime), between the 25% and 75% quantile (normal regime) or above the 75% quantile (high regime). The out-of-sample evaluation period spreads 2010:M1 to 2017:M2.

D Additional Tables

Table 7: DM tests for monthly τ_t.

	SE			
	$k = 1$	$k = 5$	$k = 10$	$k = 20$
Full sample	0.874	0.819	0.660	0.447
Low regime sample	0.650	0.490	0.298	0.177
Normal regime	0.862	0.774	0.579	0.347
High regime	0.680	0.581	0.380	0.234
	QLIKE			
Full sample	0.925	0.865	0.744	0.526
Low regime sample	0.659	0.483	0.307	0.186
Normal regime	0.872	0.780	0.579	0.363
High regime	0.788	0.698	0.482	0.294

Notes: The numbers are empirical rejection frequencies of the Diebold-Mariano test for equal predictive accuracy in favor of the M-GARCH model at the nominal 5% level. k denotes the forecast horizon. Diebold-Mariano tests in the low- and high-volatility regime are based on 60 non-overlapping observations and on 120 observations in the normal regime. The tests are based on 2000 Monte-Carlo replications.

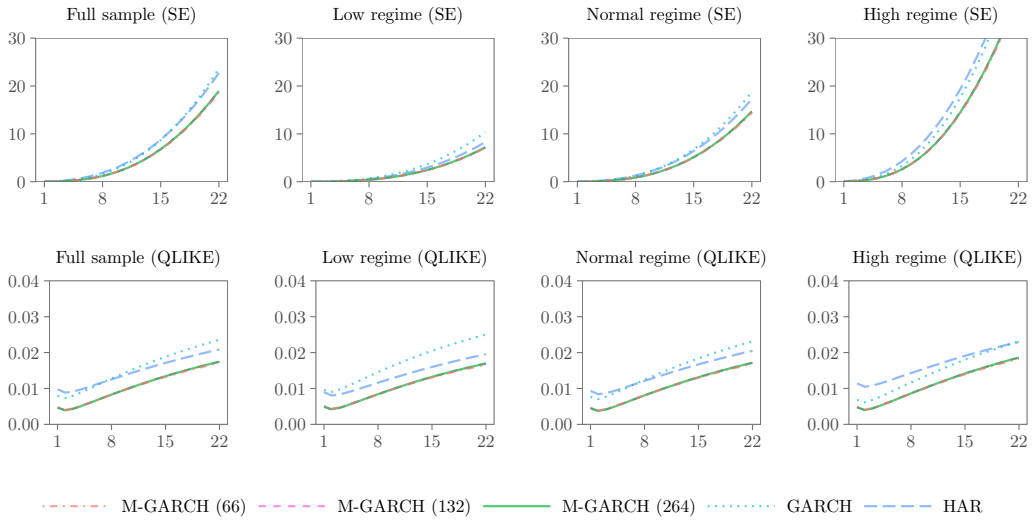
Table 8: DM tests for daily τ_t (low VR).

	SE			
	$k = 1$	$k = 5$	$k = 10$	$k = 20$
Full sample	0.922	0.868	0.745	0.536
Low regime sample	0.737	0.618	0.413	0.278
Normal regime	0.920	0.844	0.673	0.463
High regime	0.696	0.618	0.413	0.240
	QLIKE			
Full sample	0.977	0.946	0.872	0.694
Low regime sample	0.744	0.627	0.424	0.290
Normal regime	0.939	0.877	0.709	0.518
High regime	0.834	0.753	0.540	0.338

Notes: The numbers are empirical rejection frequencies of the Diebold-Mariano test for equal predictive accuracy in favor of the M-GARCH model at the nominal 5% level. k denotes the forecast horizon. Diebold-Mariano tests in the low- and high-volatility regime are based on 60 non-overlapping observations and on 120 observations in the normal regime. The tests are based on 2000 Monte-Carlo replications.

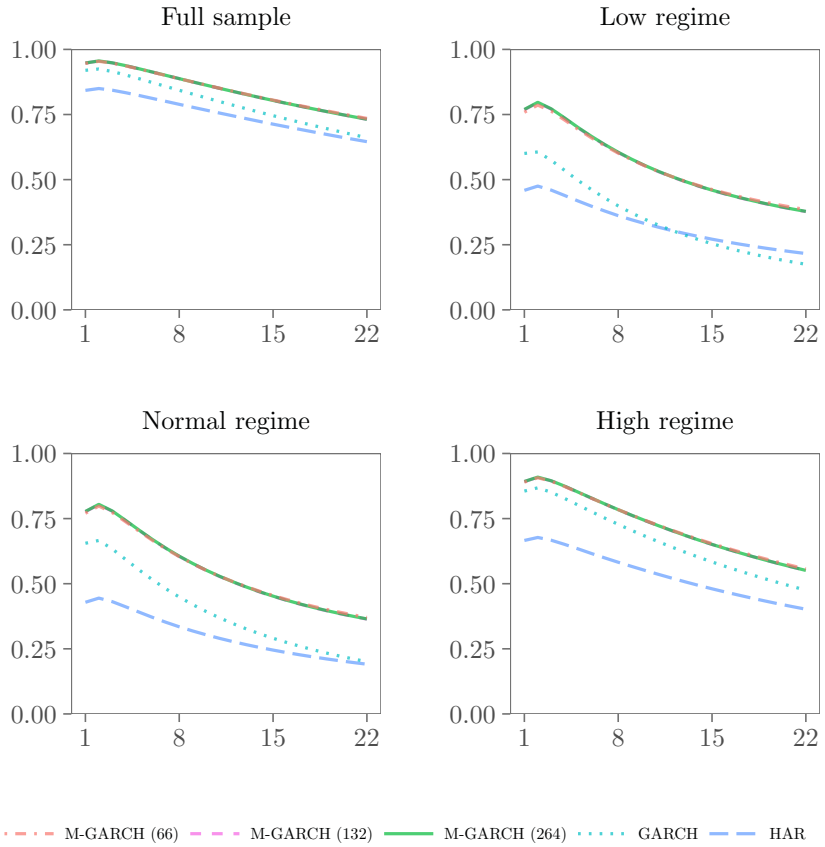
E Additional Figures

Figure 15: Daily τ_t (low VR) - SE and QLIKE loss across regimes.



Notes: Average SE and QLIKE loss evaluated at different regimes in the case of a daily-varying τ_t with a low variance ratio. In each simulation, we calculate out-of-sample SE and QLIKE loss as well as losses depending on three volatility regimes for 1–22 days ahead; forecasts being issued at a day for which the daily realized volatility is below the empirical 25% quantile (low regime), between the 25% and 75% quantile (normal regime) or above the 75% quantile (high regime). Averages are taken across 2000 simulations.

Figure 16: Mincer-Zarnowitz R^2 - daily τ_t (low VR).



Notes: Average coefficients of determination of Minzer-Zarnowitz regressions for different volatility regimes. Hereby, we regress the cumulative realized variance on the cumulative forecast. The panel displays the case of a daily-varying long-term component. In each simulation, we calculate out-of-sample R^2s depending on three volatility regimes; forecasts being issued at a day for which the daily realized volatility is below the empirical 25% quantile (low regime), between the 25% and 75% quantile (normal regime) or above the 75% quantile (high regime). Averages are taken across 2000 simulations.

F Data

In this section, we provide detailed information on the data sources as well as on the data vintages that have been used. Whenever possible, we use real-time vintage data sets as available in ALFRED.[26] For downloading the respective data sources, we have written the R-package *alfred* (Kleen, 2017).[27] We make use of the following time series:

- Realized volatility based on 5-minutes intra-day returns which are provided by the Realized Library of the Oxford-Man Institute of Quantitative Finance (Heber et al., 2009).[28]
- The VIX index as a measure of option-implied volatility of S&P 500 returns (published by the Chicago Board Options Exchange).[29]
- The Chicago Fed's National Financial Conditions Index (NFCI),[30] measuring the risk, liquidity and leverage of money markets, debt and equity markets, and the traditional and shadow banking system. The NFCI takes positive/negative values whenever financial conditions are tighter/looser than on average.
- The Chicago Fed National Activity Index (NAI) is a weighted average of 85 filtered and standardized economic indicators.[31] Whereas positive NAI values indicate an expanding US-economy above its historical trend rate, negative values indicate the opposite.
- Industrial Production Index (IP), which is released by the Board of Governors of the Federal Reserve System.[32]
- New Privately Owned Housing Units Started (HOUST), which is published by the U.S. Bureau of the Census.[33]

For the macroeconomic variables, we report the real-time data availability in Table 9. Estimations have been carried out using QMLE, see Engle et al. (2013), and can be replicated using the R-package *mfGARCH* (Kleen, 2018).[34] The covariates are depicted in Figure 7.

[26] https://alfred.stlouisfed.org
[27] https://cran.r-project.org/package=alfred
[28] http://realized.oxford-man.ox.ac.uk/data/download/
[29] http://www.cboe.com/micro/vix/historical.aspx
[30] https://alfred.stlouisfed.org/series?seid=NFCI
[31] https://alfred.stlouisfed.org/series?seid=CFNAI
[32] https://alfred.stlouisfed.org/series?seid=INDPRO
[33] https://alfred.stlouisfed.org/series?seid=HOUST
[34] https://cran.r-project.org/package=mfGARCH

Table 9: Real-time data availability.

Varibale	Frequency	ALFRED ID	First Vintage Release
NFCI	weekly	NFCI	2011-05-25
NAI	monthly	CFNAI	2011-05-23
Industrial production	monthly	INDPRO	1973-12-14
Housing starts	monthly	HOUST	1973-12-18

Notes: For each macroeconomic variable, we report the real-time data availability in the ALFRED data base.