Comparing volatility models at multiple horizons

Kevin Sheppard[*]

kevin.sheppard@economics.ox.ac.uk

Department of Economics

University of Oxford

October 6, 2015

Abstract

This paper introduces studies the performance of common volatility forecasting models across alternative forecast generation schemes that include both iterative and direct forecasts. Iterative models are found to out-perform direct forecasting methods across a wide range of horizons and assets.

Keywords: direct forecasting, forecast evaluation, GARCH, MIDAS, realized variance, volatility forecasting,

JEL Codes: C22, C53, G32

[*] The views expressed in this paper are those of the author and no other organization.

1 Introduction

This introduction is just for Asger. There will be a better one by the end of the day tomorrow, but I didn't want to screw you over any more. This paper does the following:

- Defines a direct forecast from a GARCH model
- Contrasts this with both forecast from indirect and "aggregated" GARCH models, where the aggregated GARCH model is one which uses standard low frequency data to forecast. For example, when forecasting volatility over 22 day use the 22 day return.
- Describes two methods to estimate direct forecasts, one which uses returns (like a standard GARCH) and one which uses realized-variance-type shocks (like a multiplicative error model).
- Uses a set of 25 asset return series spanning equities, bonds, foreign exchange and commodities to study the performance of these alternatives
- Also studies the effect of using longer or shorter windows
- The conclusion is better written and tells you what I found, so you might want to start there
- I hope is the a very easy paper to discuss :-)
- There will be more tomorrow, and I'll send updates as they roll off the presses. I won't hold you to any expectation of discussing them.
- It is probably somewhat badly written, but I'm working on it
- Looking forward to Brazil

2 Volatility Models

Standard GARCH models describe volatility dynamics using past returns

$$\begin{aligned} r_t &= \mu_t + \epsilon_t \\ \sigma_t^2 &= \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \\ \epsilon_t &= \sigma_t e_t \\ e &\sim F(0,1), \end{aligned}$$

where F is some known distribution – often assumed to be the Normal – with mean 0 and variance 1 and the conditional mean, μ_t has been left unspecified. Forecasting at any horizon from a GARCH model is straight-forward using the simple recursion

$$\begin{aligned} \mathrm{E}_t\left[\sigma_{t+1}^2\right] &= \omega + \alpha\epsilon_t^2 + \beta\sigma_t^2 \\ \mathrm{E}_t\left[\sigma_{t+h}^2\right] &= \omega + (\alpha+\beta)\mathrm{E}_t\left[\sigma_{t+h-1}^2\right] \end{aligned}$$

so that the cumulative h—period variance is then

$$\mathrm{E}_t\left[\sum_{j=1}^{h}\sigma_{t+j}^2\right] = \omega\left\{\sum_{j=0}^{h-1}(h-j)\theta^j\right\} + \left\{\sum_{k=0}^{h-1}\theta^k\right\}\left(\alpha\epsilon_t^2 + \beta\sigma_t^2\right) \tag{1}$$

where $\theta = \alpha + \beta$.

GARCH models with more complex lag structures can also be considered, and the component GARCH of Engle & Lee (1999) is a particularly interesting case.

$$\begin{aligned} \sigma_t^2 &= \sigma_{s,t}^2 + \sigma_{l,t}^2 \\ \sigma_{s,t}^2 &= \alpha\left(\epsilon_{t-1}^2 - \sigma_{l,t-1}^2\right) + \beta\sigma_{s,t-1} \\ \sigma_{l,t} &= \omega + \rho\sigma_{l,t-1} + \phi\left(\epsilon_{t-1}^2 - \sigma_{t-1}^2\right) \end{aligned}$$

where all parameters are positive, $\alpha+\beta<\rho<1$ and $\phi<\beta$. $\sigma^2_{l,t}$ is more persistent than $\sigma^2_{s,t}$ and so is known as the long run (or trend) component, while $\sigma^2_{s,t}$ is the short run (or cyclic) component. The component GARCH model is a restricted GARCH(2,2),

$$\begin{aligned}\sigma^2_t &= \omega\left(1-\alpha-\beta\right)+\left(\phi+\alpha\right)\epsilon^2_{t-1}-\left(\phi\left(\alpha+\beta\right)+\alpha\rho\right)\epsilon_{t-2}+\left(\rho+\beta-\phi\right)\sigma^2_{t-1}+\left(\phi\left(\alpha+\beta\right)-\rho\beta\right)\sigma^2_{t-2}\\ &= \tilde{\omega}+\tilde{\alpha}_1\epsilon^2_{t-1}+\tilde{\alpha}_2\epsilon^2_{t-2}+\tilde{\beta}_1\sigma^2_{t-1}+\tilde{\beta}_2\sigma^2_{t-2}\end{aligned}$$

and so can easily be use to produce multi-period forecasts.

An increasingly popular model is the Heterogeneous Autoregressive model of Corsi (2009) which is a highly restricted, long-lag ARCH model. The typical specification uses 1, 5 and 22-day averages of volatility shocks to model the conditional variance,

$$\begin{aligned}\sigma^2_t &= \omega+\alpha_1\epsilon^2_{t-1}+\alpha_5\left({}^1\!/_5\sum_{j=1}^{5}\epsilon^2_{t-j}\right)+\alpha_{22}\left({}^1\!/_{22}\sum_{k=1}^{22}\epsilon^2_{t-j}\right).\\ &= \omega+\sum_{j=1}^{22}\phi_j\epsilon_{t-j},\end{aligned}$$

where $\phi_j=\alpha_1+{}^{\alpha_5}\!/_5+{}^{\alpha_{22}}\!/_{22}$ when $j=1$, ${}^{\alpha_5}\!/_5+{}^{\alpha_{22}}\!/_{22}$ for $j\in[2,5]$ and ${}^{\alpha_{22}}\!/_{22}$ for $j\in[6,22]$. Since this is just a restricted ARCH model, multi-step forecasts follow the usual recursive form

$$\mathrm{E}_t\left[\sigma^2_{t+h}\right]=\omega+\sum_{j=1}^{22}\phi_j\mathrm{E}_t\left[\sigma^2_{t+h-j}\right].$$

Both the common 22-lag HAR and a "quarterly" HAR which also includes a 66-day lag will be included.

MIDAS volatility models Ghysels, Santa-Clara & Valkanov (2002) provide an alternative method to parameterize long-lag ARCH-type models. MIDAS volatility models use parameterized weight functions to limit the number of that must be estimated while providing more flexibility in the manner weights decay than in standard GARCH models. The generic form of the conditional variance in a MIDAS model is

$$\sigma^2_t=\omega+\alpha\sum_{j=1}^{m}\phi_j(\theta)\epsilon^2_{t-j}.$$

The weighting function, $\phi_j(\theta)$ is non-negative, sums to 1 and is determined by a small-dimensional parameter vector θ; m is the maximum lag used. When θ is known (or can be estimated) then the structure of a MIDAS volatility model is identical to that of a HAR, only using a different set of weights. A number of weighting functions have been considered including ones base off of exponential functions, the Beta distribution, and gamma functions. The latter produces weights with hyperbolic-like decay.

Here I consider specifications. The first is the Beta distribution-based weighting function,

$$\phi_j(\theta) \propto \frac{(j/m)^{\theta_1-1}\left(1-(j/m)^{\theta_2-1}\right)}{\Gamma(\theta_1)\Gamma(\theta_2)/\Gamma(\theta_1+\theta_2)}$$

which depends on a two-dimensional positive parameter θ. The second is the the single parameter hypergeometric version,

$$\phi_j(\theta) \propto \frac{\Gamma\left(j+\theta\right)}{\Gamma\left(j+1\right)\Gamma\left(j+\theta\right)},$$

where $\theta > 0$. The final is the exponential-based weighting function,

$$\phi_j(\theta) \propto \exp\left(\theta_1 j + \theta_2 j^2\right),$$

where $\theta_2 < 0$ ensures that long-lag weights will converge to 0. Both the Beta distribution-based and the hypergeometric weight function are truncated at a finite number of lags, m. Two version of these specifications will be used, one with $m = 22$ and the other with $m = 66$.

2.1 Estimation

Estimation of the models used to produce iterative forecasts will be performed using Gaussian QMLE. Returns, conditional on variance, are assumed to be normally distributed, $r_{t+1}|\mathscr{F}_t \sim N\left(\mu_t, \sigma_t^2\right)$. In all specification $\mu_t = \mu \; \forall \; t$ is assumed , since the daily mean is extremely small relative to the daily standard deviation of returns (Andersen & Bollerslev 1998). The iterative forecasts are all produced by estimating parameters using the 1-day ahead conditional quasi likelihood,

$$\max_{\theta}\sum_{t=1}^{T}\ln\sigma_t^2+\frac{\epsilon_t^2}{\sigma_t^2}.$$

2.2 Direct Estimation

The multi-step forecast from a GARCH model shows that it only depends on model parameters and the final values of ϵ_t^2 and σ_t^2. This leads to a natural questions as to whether the model-imposed parameter restrictions are valid, especially over longer horizons. A direct version of a GARCH(1,1) can be specified as

$$\mathrm{E}_t\left[\sum_{j=1}^{h}\sigma_{t+j}^2\right]=\omega_h+\alpha_h\epsilon_t^2+\beta_h\sigma_t^2 \tag{2}$$

which nests the iterated forecast by matching ω_h, α_h and β_h to their counterparts in 1. Throughout $\sigma_{t+i:h}^2\equiv\sum_{j=1}^{h}\sigma_{t+j}^2$ will be used to denote the h-period variance. Direct estimation of a model of this type requires using a horizon specific quasi-likelihood,

$$\max_{\theta_h}\sum_{t=1}^{T}l_t=\max_{\theta_h}\sum_{t=1}^{T}\ln\sigma_{t+1:h}^2+\frac{\left(\sum_{j=1}^{h}\epsilon_{t+j}\right)^2}{\sigma_{t+1:h}^2}, \tag{3}$$

where θ_h is used to acknowledge that for a particular model, the parameters will depend on the horizon used in the estimation.

When models, are correctly specified at the 1-step horizon, direct estimates of the model parameters are also consistent for the same values. They will, however, be inefficient since fitting the h-day squared innovation add unnecessarily noise to the measurement.

Proposition 1. *Let the model for the conditional variance be given by a GARCH(1,1) and further assume the model is dynamically correct in the sense that $\sigma_t^2=E_{t-1}\left[\epsilon_t^2\right]$ where the true parameters are θ_0. Then the expectation of the score of the log-likelihood of the h-period direct quasi-likelihood,*

$$\mathrm{E}\left[\frac{\partial l_t}{\partial\theta}\right]_{\theta=\theta_0}=0$$

where $\omega_h = \omega\left\{\sum_{j=0}^{h-1}\left(h-j\right)\theta^j\right\}$, $\alpha_h = \alpha\left\{\sum_{k=0}^{h-1}\theta^k\right\}$ *and* $\beta_h = \beta\left\{\sum_{k=0}^{h-1}\theta^k\right\}$.

The GARCH model is unique among the other models in the sense that the method of parameterizing a direct forecast, either using the simple specification

$$\sigma^2_{t+1:h} = \omega_h + \alpha_h \epsilon^2_t + \beta_h \sigma^2_t$$

or the implicit specification in terms of the 1-step ahead specification,

$$\sigma^2_{t+1:h} = \omega\left\{\sum_{j=0}^{h-1}\left(h-j\right)\theta^j\right\} + \alpha\left\{\sum_{k=0}^{h-1}\theta^k\right\}\epsilon^2_t + \beta\left\{\sum_{k=0}^{h-1}\theta^k\right\}\sigma^2_t$$

lead to identical parameter estimates. The is not generally true when there are more parameters in the ARCH specification than in the model. For example, consider a HAR model, where the 1-step ahead forecast is

$$\mathrm{E}_t\left[\sigma^2_{t+1}\right] = \omega + \alpha_1\epsilon^2_t + \alpha_5\left({}^1\!/_5\sum_{j=1}^{5}\epsilon^2_{t-j+1}\right) + \alpha_{22}\left({}^1\!/_{22}\sum_{k=1}^{22}\epsilon^2_{t-j+1}\right).$$

Since this is the natural one-step ahead forecast, there are 3 unique parameters in the dynamics. The two-step ahead forecast is

$$\mathrm{E}_t\left[\sigma^2_{t+1:2}\right] \quad = \quad \omega\left(1+\phi_1\right) + \sum_{j=1}^{21}\left(\phi_1\phi_j + \phi_{j-1}\right)\epsilon^2_{t-j+1} + \phi_1\phi_{22}\epsilon^2_{t-21}. \tag{4}$$

In terms of the original model parameters, α_1, α_5 and α_{22}, there are 5 distinct values: $\alpha_1 + \alpha_1\alpha_5/5$ for ϵ^2_t, $\alpha_1\alpha_5/5+\alpha_5/5$ for $\epsilon^2_{t-1},\ldots,\epsilon^2_{t-3}$, $\alpha_1\alpha_5/5+\alpha_{22}/22$ for ϵ^2_{t-4}, $\alpha_1\alpha_{22}/5+\alpha_{22}/22$ for $\epsilon^2_{t-4},\ldots,\epsilon^2_{t-20}$ and $\alpha_1\alpha_{22}/22$ for ϵ_{t-21}. In general the h-step cumulative forecast of a HAR model will depend on the constant and $\min(1+2h,22)$ unique parameters on the lagged squared residuals. When models are restricted ARCH-type models there will not be a direct one-to-one correspondence between the 1-step ahead model and the h-step ahead model.

It is, however, possible to enforce the correspondence by directly estimating the forward-iterated version of a model. Again, consider the example of the HAR model. Eq. 4 only depends on the original constant and α_1, α_5 and α_{22}, and so direct estimation of a model that is compatible with the one-step-ahead model is trivial. This is true for any model which has an ARCH representation by specifying the parameters in terms of the one-step-ahead specification irrespective of the estimation frequency. In this version of the paper, all direct forecasts are produced using the simple approach so that

$$\sigma^2_{t+1:m} = \omega + \sum_{i=1}^{\infty} \phi_i(\theta)\epsilon^2_{t-i+1}.$$

2.3 Aggregated GARCH

This definition of a direct forecast differs from Ghysels, Rubia & Valkanov (2009) who define the direct model as a standard GARCH using aggregated data with horizon h. Their version of a direct model is then

$$\mathrm{E}_t\left[\sum_{j=1}^{h}\sigma^2_{t+j}\right] \approx \tilde{\omega}_h + \tilde{\alpha}_h\left(\sum_{j=0}^{h-1}\epsilon_{t-j}\right)^2 + \tilde{\beta}_h\tilde{\sigma}^2_t \tag{5}$$

where the approximation arises since the conditional mean would generally differ. This version just a standard GARCH model based on lower-frequency data. This definition is not, however, consistent with what is commonly done in the macroeconometric literature, see e.g., Marcellino, Stock & Watson (2006), and imposes artificial limits on the lag structure when compared to the natural direct model in eq. (2). I will refer the model in eq. 5 as the aggregated forecasting model, since if returns have mean 0 and are serially uncorrelated,

$$\mathrm{E}_t\left(\sum_{j=1}^{h}\epsilon_{t+j}\right)^2 = \mathrm{E}_t\left(\sum_{j=1}^{h}\epsilon^2_{t+j}\right)$$

and $\mathrm{V}\left[\left(\sum_{j=1}^{h}\epsilon_{t+j}\right)^2\right] > \mathrm{V}\left[\sum_{j=1}^{h}\epsilon^2_{t+j}\right]$ so that

$$\mathrm{E}_t\left[\sum_{j=1}^{h}\sigma^2_{t+j}\right] \approx \tilde{\omega}_h + \tilde{\alpha}_h\sum_{j=0}^{h-1}\epsilon^2_{t-j} + \tilde{\alpha}_h\eta_t + \tilde{\beta}_h\tilde{\sigma}^2_t$$

where η_t is a noise term that arises since the squared sum is a much noisier estimator than than the sum of squared returns. This version of the aggregated model shows that aggregating does two things: first, it imposes a very flat lag structure on volatility innovations and second it adds noise. The lag structure can be flexibly addressed in the direct equation, and there is little reason to believe including extra noise in the volatility innovation measurement would improve the forecast. I also consider a modified version of the aggregated model using a $h-$period innovation using the ideas from realized variance on the right hand side in place of the squared h-period return (Andersen et al. (2001), Barndorff-Nielsen & Shephard (2002)). The volatility dynamics in this model

$$\sigma^2_{t+1:h} = \omega_h + \alpha \sum_{i=1}^{h} \epsilon^2_{t-1+1} + \beta \sigma^2_{t-h+1:0}.$$

This specification is just a standard h-period GARCH(1,1) where the usual squared return innovation has been replaced by the $h-$period realized variance.

2.4 Low-frequency Models

Models based on the one-step ahead fit are estimated using all data.[1] Models that are based on longer horizon returns, such as the direct forecasts or the aggregated models would usually only be estimated using a subset – every h^{th} observation – and so do not efficiently utilize the available information. This also adds a degree of arbitrariness since there are h initial observations that can be used to estimate the model. To address this, I used an overlapping block estimators. Given a parameter vector θ, these estimators make use of h recursions to compute h quasi log-likelihoods. The estimated parameter maximizes the sum of these h-log likelihoods.

2.5 Naïve Models

Historical volatility is popular among practitioners since it is both simple to compute and interpret. The only parameter required to operationalize historical variance is the window length, and three horizons

[1] Any backcast values are set to the same value σ_0^2 which is based on an exponentially weighted moving average $\sigma_0^2 = (1-\lambda)\sum_{t=1}^{T} \lambda^t \epsilon_t^2$. All estimates make use of all data irrespective of the lag structure.

will be considered: monthly (22 days), quarterly (66) and annual (252). The estimator is defined

$$\sigma_t^2 = \sum_{i=1}^{k} \left(r_{t-i} - \mu\right)^2$$

where k is one of the window lengths.

3 Data

The models are evaluated using a range of assets spanning the major asset classes: equities, treasuries, corporate bonds, currencies and commodities. The equity series include the value weighted market (VWM), the value factor (HML, Fama & French (1992)), and a momentum factor (UMD, Carhart (1997)), as well as the components of 12 industry portfolios. These were all taken from Ken French's return data library.[2] Returns to treasury portfolios were constructed to correspond to the return on level, slope and curvature portfolios. These portfolios were constructed using the 1, 5 and 10 year treasury yields. Yields were transformed into prices which were then transformed into log returns. The level is the holding return on the 1-year bond, the slope is the return on a portfolio the is long the 10-year bond and short the 1-year bond, and the curvature factor is the return on a portfolio that is long both the 1 and 10 year bonds and short the 5 year bond. The currency returns are from the USD-GBP exchange rate, the USD-JPY exchange rate and a trade-weighted US Dollar index. The BoA Merrill Lynch US Corp Master Total Return Index was used to measure the return on a corporate bond portfolio. Commodities are represented by gold, west Texas intermediate crude, and the CRB Commodity Index. All non-equity data series were taken from from the Federal Reserve Economics Database, with the exception of the CRB commodity index which comes from Haver. Table 1 contains additional details on the source of data used. All series were included from the earliest date that daily data was continuously available until the end of December 2014.

The range of series was chosen to provide a wide range of data characteristics and especially to go beyond the common benchmarks of a broad, value-weighted equity portfolio. Summary statistics of the

[2] The series are available at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

data are presented in 2. The panel has a wide range of basic characteristics, with annualized volatilities ranging from the low single digits to nearly 40% for crude oil. The skewness values range from -1.6 for the momentum portfolio for 0.5 for the Treasury level factor portfolio. The kurtosis of the data series also spans a wide range, from only slightly higher than that of a normal for corporate bonds returns to over 30 for the momentum portfolio.

3.1 In-sample Results

Before turning to forecast evaluation, it is useful to assess the "best case" scenario for these model using some in-sample measures of fit. Table 3 contains relative log likelihood values and average ranks. Given the range of the sample sizes across the the 24 series, a direct comparison of log-likelihood differences would be strongly biased towards the series with more data points. To remedy this, all values in the table are based the average daily log-likelihood value scaled up to a typical hypothetical 10-year sample with 2,520 days. Using these scaled log-likelihoods, the difference relative to the median was computed series-by-series. The average of these differences is reported in the left panel.

The best model, on average, was the component GARCH model which typically outperformed the median model by 15 or more log-likelihood points across all series. Other models which performed well include all models that feature use 66 lags: the HAR$_{66}$, the MIDAS$-\beta_{66}$ and the MIDAS-hyp$_{66}$. Models with shorter lag lengths or with less flexibility performed worse than the median in the majority of series.

The two aggregated GARCH models performed very differently. Results for these two models are only presented for the multi-step horizons since they exactly coincide with a GARCH model when $h=1$. The standard aggregated GARCH model which makes use of h-day returns as the innovation was much worst than the median model. The realized-variance-based version of the aggregated GARCH model performs much better, and is typically very close to the median model. This shows the gains to using the sum of squares rather than the square of the sum when modeling longer horizon returns. The worst models are clearly the moving average models, which is not surprising since these have no free parameters to optimize.

The ranks paint a similar picture, with the component GARCH model performance the best for nearly all series. The three other models that performed well on average also have similar, low ranks, and models with shorter lag lengths are worst than average. The return-based Aggregate GARCH model is only better than the moving averages while the RV-based aggregated GARCH model performs slightly better than average for all horizons.

Figure 1 contains the first 30 weights from an ARCH representation for the GARCH and Component GARCH models and the 22- and 66-lag versions of the MIDAS$-\beta$ and MIDAS-hyp models using data on the value weighted market portfolio and the entire sample. The longer lag versions consistently outperformed the shorter lag models, even when the models were not strictly nested. A common feature of the longer lag models is the hyper-geometric-like shape where a relatively large weight is given to the most recent observation followed by a rapid decline in weights for intermediate values followed by a relatively flat set of weights given to long lags.

4 Forecast Comparison

Forecasts were generated for all models using a variety of configurations. The baseline estimation method used a 10-year rolling window scheme and all quasi-likelihoods were computed using returns (eq. 3). Other parameterizations made use of 5- or 20-year rolling windows or altered the quasi-log likelihood to use realized variance (eq. XX). All models were re-estimated monthly and the parameters were held fixed until the start of the next calendar month.

4.1 Forecast Evaluation

The forecast evaluation focuses on relative performance of the alternative models. Forecasts are assessed using the QLIK loss function,

$$L\left(\hat{\sigma}^2_{t+1:h}, r_{t+1}, \ldots r_{t+h}\right) = \ln \hat{\sigma}^2_{t+1:h} + \textstyle\sum_{i=1}^{h} r_{t+i}/\hat{\sigma}^2_{t+1:h}.$$

Volatility forecast evaluation is sensitive to the choice of loss function since the object of interest, the conditional variance, is not adapted to the $\mathscr{F}_{t+h}$ in formation set. The set of loss functions that will prefer the true model even when using a noisy proxy has been studiesHansen & Lunde (2005) and Patton (2011). The QLIK loss function is in this set, as is the MSE loss function. Simulations in Patton & Sheppard (2009) show that the QLIK version performs better than other parameterizations, and so this version is used here.

When comparing two forecasts, Diebold-Mariaon-Giaomini-White test statistics are a natural choice (Diebold & Mariano (1995), Giacomini & White (2006)). The DMGW test statistic is

$$\sqrt{P}\left(\sum_{t=R+1}^{T} L\left(\hat{\sigma}^2_{t+1:h,A}, r_{t+1}, \ldots r_{t+h}\right) - L\left(\hat{\sigma}^2_{t+1:h,B}, r_{t+1}, \ldots r_{t+h}\right)\right) = \sqrt{P}\bar{\delta}.$$

where P is the number of out-of-sample observations and A and B are used to represent the forecasts of two competing models. In practice, the test statistic is the same in either Diebold & Mariano (1995) or Giacomini & White (2006), and the main difference in in the assumptions. Giacomini & White (2006) incorporate estimation error into the loss function so that two models, even when nested, can be compared since the amount of parameter estimation error will differ. Under standard assumptions, even when models are nested,

$$\sqrt{P}\bar{\delta} \xrightarrow{d} N(0, V)$$

where V is the long-run variance of the difference. In practice a Newey & West (1987) estimator is used with a data driven bandwidth. The null of the test is equal predictive accuracy, $H_0 : \mathrm{E}[\delta_t] = 0$ and the composite alternative is $H_1^A : \mathrm{E}[\delta_t] < 0$ and $H_1^B : \mathrm{E}[\delta_t] > 0$. The forecasts were produced using a rolling estimation scheme and so incorporating the consequences of parameter estimation error is desirable.

Pairwise DMGW tests, while useful, are difficult to extend to multiple models. The Model Confidence Set (MCS) of **?** provides a method to extend the pairwise tests to a collection of tests. The MCS aims to find the set of models which are not distinguishable from the best model while controlling the familywise error rate – that is the probability that the best models (or a model that is equal good in predictive accuracy as the best) is excluded. Conceptually this is similar to the size of the test. The MCS is constructed by examining all pairwise differences in losses and estimating the distribution of these differences if all models were equally accurate. Models which are unlikely to to be as good as the best model – that is, those with losses that would have been unlikely to be seen if they were equally good are excluded, and the algorithm is rerun on the remaining models. This algorithm provides a sequence of p-values where each model was excluded. The final MCS is all models that could not be excluded at the α level. In all applications of the MCS, α =10% is used.

4.2 Results

Each series had a total of 324 distinct forecasts spanning the range of:

- 11 dynamics models plus 3 moving averages[3]
- 4 forecast horizons – 1, 5, 10 and 22 days
- 3 forecasting methods – iterative, direct using return-based estimates and direct using realized variance-based estimates
- 3 estimation windows – 5, 10 and 20 years

The results will be presented sequentially covering one feature at a time.

Ranking Iterative Models

Table 4 contains results for pairwise DMGW tests for the iterative forecasts across the 4 horizons. Each statistic was computed for all series and all combinations of models. Test statistics larger than 1.96 indicated that model B – the column model – outperformed model A – the row models. The value reported is the percentage of rejections across the 25 series for each combination.

For horizon 1 – the estimation horizon for all models – there is a clear set of models that perform relatively well and a clear set that perform relatively poorly. The best performing model is the component GARCH, which is only worse than two other models for a small number of series. The component GARCH model strongly outperforms models with shorter memory including the standard GARCH, and 22-lag versions of the HAR, MIDAS$-\beta$, MIDAS$-$exp and MIDAS-hyp. The other models that outperform more than they under perform include the 66-lag version of the HAR, MIDAS$-\beta$ and MIDAS-hyp.

[3] Not all models are unique across all configurations. The moving average models all produce the same forecasts for any forecasting method or horizon and so do not vary across configurations. The aggregated GARCH models do not differ from standard GARCH when estimated at horizon 1 and so these results are excluded, and are only used for direct forecasts. Finally, when the horizon is 1, all methods are identical for all models.

The worst model is the MIDAS$-$exp which is statistically worse then the alternative in over 50% of the comparisons and rarely outperforms other models. Other lower persistence models including the standard GARCH and the 22-lag versions of the HAR, MIDAS-β and MIDAS-hyp all under perform.

Increasing the horizon allows the two aggregated GARCH models to be included in the comparison. There is a clear patter here: across all longer horizon forecast evaluation samples, the standard aggregated GARCH which uses the h-period return as the shock is never better than any model and is almost always worse than the other models. The aggregated GARCH model that uses a realized-variance-type shock also perform relatively poorly except when compared to the other aggregated model. The increase in the horizon reaffirms and strengthens the pattern observed at the shorter horizon – a small number of models are clearly superior, and many common models perform worse than their competitors. The set of superior models clearly includes the component GARCH, and the 66-lag versions of the MIDAS-β and MIDAS-hyp models. The 66-lag version of the HAR performs relatively better than the lower persistence models, but either outperformed or no better than the set of three superior models. The lower persistence models are roughly similar with no obvious pattern aside from the consistent under performance of the MIDAS$-$exp model.

Pairwise comparisons are difficult to interpret since controlling size is challenging when the test statistics are likely to be dependent. The MCS was applied to the set of models at each horizon on a series-by-series basis. Table 5 contains the percentage of Model Confidence Sets where a model appears using a family wise error rate of 10%. The left panel applies the MCS on the losses directly. Three models stand out with inclusion rates above 90% across the horizons: the component GARCH and the 66-lag versions of the MIDAS-β and MIDAS-hyp. The 66-lag HAR is also in the majority of MCS. The lower persistence models – GARCH, HAR$_{22}$, MIDAS$-\beta_{22}$ and MIDAS-hyp$_{22}$ are included a lower rate, although they still appear in nearly 50% of the sets. The MIDAS-exp is notable worse, and both aggregated GARCH models and the moving average models are excluded from most sets.

The right panel of table 5 applies the MCS to the ranks of the losses. For each series, for each forecast observation, the losses were ranked across all models so that the lowest loss received a rank of 1 and the highest a rank equal to the number of models. While the rank of the QLIK loss function is not necessarily robust to noise in the proxy, using ranks in place of losses can be considered a robustness measure to ascertain whether the performance is driven by the typical loss or by a relatively small number of observations with large losses. There is a strong correspondence between the rank MCS and the loss MCS with one notable exception. The 22-day moving average performs exceptionally well in terms of ranks at short horizons and is in 100% of the MCS. At longer horizons the MCS tend to include the component GARCH and the 66-lag models, with the HAR_{66} in all but 2 of the MCS when evaluated at a 22-day horizon.

In addition to directly exampling the MCS on the series, both losses were aggregated across asset-type groups as well as across all assets. Two types of aggregation was used. The first simply averages the losses across group members and the second uses a GLS inspired average where

$$\bar{L}_{M_i} = \frac{\sum_{j \in Group} w_j L_{j,M_i}}{\sum_{j \in Group} w_j}$$

where the weights were computed as the inverse of the average (across models) variance of the losses. This GLS-type average gives more weights to series that have relatively less time series variation in their loss, which typically comes from having less volatile volatility. Table 6 contains the average inclusion rates across the four groups: equities, commodities, interest rates and exchange rates. Grouping broadly confirms that the component GARCH as well as the the 66-lag MIDAS models perform well and appear in all MCS, irrespective of the averaging method. The only model that performs differently across the aggregation methods is the GARCH which is in most MCS when using the GLS series. This indicates that GARCH performs relatively well in series that have less volatile volatility. The third panel in table 6 contains results using the average rank in place of the average loss. Since ranks are unlikely be highly heterostructure, only a simple average is included. The rank-based results are similar to the previous rank-based finding where the moving average performs much better especially at short horizons, and the lower persistence models are excluded from all of the MCS.

Finally, table 7 contains the MCS p-values for the fully aggregated losses across all 25 series. The left panel contains the results from a simple average and the middle panel contains the p-values for a GLS average of the losses and the right panel contains the MCS p-values for the average ranks. These results are consistent with what has been found both for the individual assets and the grouped averages. Three models standout – the component GARCH, and the 66-lag versions of the MIDAS–β and the MIDAS-hyp. Using a FWER of 10%, the only other model that would be included is the 66-lag version of the HAR. All of the lower persistence models and the moving averages are excluded, typically with very small p-values. Using GLS weighting makes little difference to the MCS, and the same models are included with the exception of the 66-lag HAR which is excluded for 3 of the 4 horizons. The rank-based MCS shows a similar pattern with the usual exception that a short moving average performs well for the shorter horizons.

Ranking Direct Forecasting Models

Direct forecasts were produced using two estimation schemes: the first estimates model parameters using a quasi-likelihood for h-day squared returns and the second uses a quasi-likelihood for h-day the sum of squared returns. These two methods will be referred to as the returns-base forecasts and the realized variance-based forecasts. Table 8 contains the results for the MCS constructed on average losses, GLS-averaged losses, and average ranks for the alternative direct forecasting schemes. The MCS were constructed separately for the two methods of forecasting model estimation. These tables are similar in structure to table 7.

The top panels contain the MCS p-values of models when considering only return-based forecasting models. The bottom contains results for MCS for models estimated using realized variance. When estimating models using returns a number of models consistently perform well: the component GARCH and the 66-lags versions of the HAR. The two MIDAS models that performed well in the iterative forecasting are excluded for the longer horizons. When estimating the model parameters using realized variance, the model confidence set is slightly larger and includes the component GARCH model as well as the 66-lag versions of the MIDAS-β and MIDAS-hyp models.

While these two tables show that a similar set of models performs well irrespective of the estimation or forecasting method, they do not address the issue of which models perform best across alternative forecasting generation schemes. To address this question pairwise DMGW tests were estimated comparing the iterative forecast for a particular model with the iterative forecast for both direct forecast parameter estimation methods. Table 11 contains results of these comparisons where a the test is implemented using a 5% size. When comparing iterative models to direct forecasts generated using parameters estimated on returns there is unequivocal evidence that iterative models are superior. The direct forecast was never preferred to the iterative forecast, and in most instances over 90% of the iterative forecasts were superior to the direct forecasts. When the estimation method is changes to utilize realized variance the picture is slightly altered, although the iterative models continue to perform as well or better than the direct models in most cases. The only models which show meaningful rejection of the null in favor of the direct model are for those models which perform worse, on average, including the GARCH as well as the 22-lag versions of the HAR, MIDAS-β and MIDAS-hyp. The models that were consistently top performers consistently outperform their direct counterparts.

The source of the performance difference is the extra parameter estimation error that is present in the directly estimated models. The iterative models are relatively efficient, and so the only hope for the direct models is meaningful misspecification of the iterative models. However, for a reasonably well specified iterative model — one that utilizes a relatively long lag structure – any gains in bias are more than offset by a reduction in parameter estimation error. This is confirmed when comparing the performance across alternative estimation windows. All results presented thus far have utilized 10-year rolling estimates. Table 10 contains results for model-by-model comparisons of iterated forecast versus realized-variance-based direct forecasts using both 5-year and 20-year estimation windows. [4] The top panel contains results for a 5-year estimation window and shows a clear preference for iterative forecasts when the estimation sample is shorter. This is consistent with an increase in parameter estimation error especially for the direct models. When the estimation window is lengthened to 20 years the direct models perform relatively better – often outperforming the iterative models. Lengthening the sample has two consequences for the forecasting models. The obvious one is that the parameters should be estimated with more precision. The less obvious consequence is that the model is more likely to be well specified and so the gains to direct forecasting, when evaluated at pseudo-true parameters might be larger.

One final comparison was using the MCS on an initial set of models that included the best models

[4] Changing the estimation sample alters the prediction sample so that these models are evaluated over different samples than the 10-year estimates. In all cases the maximum number of out-of-sample observations was used.

across the iterative and direct estimation schemes. MCS were constructed series-by-series for 8 models: 4 forecasts produced using iterative methods and 4 produced using direct methods. Table 11 contains the percentage of MCS that contained the model with a FWER of 10%. For most model/horizon combinations the iterative version of the model is contained in the MCS at least as often as the direct version of the model. The component GARCH, the MIDAS-β_{66} and the MIDAS-hyp are in almost all MCS when using losses. When suing average ranks, the iterated models continue to outperform the direct forecasting models.

5 Extensions

Asymmetric models TBC, Conditional predictive ability TBC

6 Conclusions

This paper examines the performance of volatility forecasting models over a range of frequencies out to one month. Three classes of models are consistently found to perform the best across a range of setups, estimation methods and data series: the component GARCH, the MIDAS$-\beta$ and the MIDAS-hyp. Other models, especially models commonly used such as a HAR using 3 components with a maximum lag of 22 days, are consistently outperformed by these models. The worse models were consistently aggregated GARCH models which directly forecast the h-day variance using h-day shocks. It is clear that using a simple approach to forecasting low frequency volatility – a week or more – is an unwise choice.

Direct forecasts are generally unable to outperform iterative forecasts. Moreover, the direct forecasts are only able to outperform if estimated using the h-day realized variance to measure variance – using h-day squared returns as the volatility proxy is sufficiently noisy that any gains from a reduction in bias are lost to increased variance of parameters. The only setups where the forecasts from direct models outperformed forecasts generated iteratively are when the estimation window is especially long – 20 years of daily data.

Among the models that were consistently found to perform well, the MIDAS with a hypergeometric weighting function is the most parsimonious, using only 3 parameters.[5] There are still a number of remaining questions when examining the relative performance of direct forecasts. Are there times when direct forecast can out-perform iterative forecasts, and so it is possible to improve these by combining information from both? Are conditional asymmetries useful for predicting long horizon volatility? I leave these as topics for further research.

[5] The lag length used could be considered a model parameter, although this is not usually optimized.

References

Andersen, T. G. & Bollerslev, T. (1998), 'Answering the skeptics: Yes, standard volatility models do provide accurate forecasts', *International Economic Review* **39**(4), 885–905.

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2001), 'The Distribution of Realized Exchange Rate Volatility', *Journal of the American Statistical Association* **96**(453), 42–55.

Barndorff-Nielsen, O. E. & Shephard (2002), 'Econometric analysis of realized volatility and its use in estimating stochastic volatility models', *Journal Of The Royal Statistical Society Series B* **64**(2), 253–280.

Carhart, M. M. (1997), 'On persistence in mutual find performance', *Journal of Finance* **52**(1), 57–82.

Corsi, F. (2009), 'A Simple Approximate Long-Memory Model of Realized Volatility', *Journal of Financial Econometrics* **7**(2), 174–196.

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

Engle, R. F. & Lee, G. J. (1999), A permanent and transitory component model of stock return volatility, *in* 'Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W.J. Granger', Oxford University Press.

Fama, E. F. & French, K. R. (1992), 'The cross-section of expected stock returns', *Journal of Finance* **47**, 427–465.

Ghysels, E., Rubia, A. & Valkanov, R. (2009), Multi-period forecasts of volatility: Direct, iterated, and mixed-data approaches, Technical report, University of North Carolina.

Ghysels, E., Santa-Clara, P. & Valkanov, R. (2002), The midas touch: Mixed data sampling regression models. UCLA.

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

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

Marcellino, M., Stock, J. H. & Watson, M. W. (2006), ‘A comparison of direct and iterated multistep ar methods for forecasting macroeconomic time series’, *Journal of Econometrics* **135**(1–2), 499–526.

Newey, W. K. & West, K. D. (1987), ‘A simple, positive definite, heteroskedasticity and autocorrelation consistent covariance matrix’, *Econometrica* **55**(3), 703–708.

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

Patton, A. & Sheppard, K. (2009), ‘Evaluating volatility and correlation forecasts’. in T.G. Andersen, R.A. Davis, J.P. Kreib and T. Mikosch, eds, Handbook of Financial Time Series, Springer, pp. 801–838.

A Tables

Code	Name	Additional Info	First Observation
Vwm	Value Weighted Market		July 1926
Hml	High minus Low factor		July 1926
Mom	Momentum factor		November 1926
Bus	Business Equipment		July 1926
Chem	Chemicals		July 1926
Durbl	Consumer Durables		July 1926
Energy	Oil, Gas and Energy		July 1926
Health	Healthcare		July 1926
Finance	Financial Firms		July 1926
Manuf	Manufacturing Firms		July 1926
NonDurbl	Consumer Non-durables		July 1926
Other	Other Firms		July 1926
Shops	Wholesale, Retails and Services		July 1926
Tele	Telephone and Television		July 1926
Utility	Utilities		July 1926
Level	1-year Yield	DGS1	June 1961
Slope	15-year Yield	DGS5	August 1971
Curve	10-year Yield	DGS10	August 1971
JPYUSD	Yes-Dollar rate	DEXJPUS	January 1971
GBPUSD	Pound-Dollar rate	DEXUSUK	January 1971
TWUSD	Trade Weighted Dollar index	DTWEXM	January 1973
Corp Bond	BoA Merrill Lynch US Corp TRI	BAMLCC0A0CMTRIV	November 1986
Gold	London-fixing Gold	GOLDAMGBD228NLBM	April 1968
Crude	West Texas Intermediate Crude	DCOILWTICO	January 1986
Comm	CRB Commodity Index	PZALL	January 1971

Table 1: Description and source of data used in paper.

	Observations	μ	σ	Skewness	Kurtosis
Vwm	23385	7.25	17.00	-0.110	19.82
Mom	23284	6.88	11.78	-1.673	31.58
Hml	23385	4.15	9.15	0.660	18.67
Bus	23385	12.40	24.39	0.205	15.90
Chem	23385	11.73	18.32	-0.241	25.19
Durbl	23385	12.19	24.20	0.371	18.45
Energy	23385	12.18	20.27	0.098	17.84
Health	23385	12.20	17.46	-0.275	19.66
Manuf	23385	11.40	20.18	0.207	23.67
Finance	23385	10.93	20.88	0.209	25.70
NonDurbl	23385	11.03	13.76	-0.237	23.41
Other	23385	9.10	19.53	-0.010	16.83
Shops	23385	11.13	17.36	-0.018	17.71
Tele	23385	9.99	16.34	0.207	21.26
Utility	23385	9.86	17.42	0.326	27.33
Level	13349	0.05	0.95	0.505	23.35
Slope	10813	1.27	9.33	0.129	8.86
Curve	10813	0.23	4.09	-0.093	26.95
JPYUSD	11036	-1.97	10.28	-0.565	12.01
GBPUSD	11042	-0.54	9.42	-0.143	7.82
TWUSD	10524	-0.35	6.71	-0.291	8.83
Corp Bond	7302	7.11	4.76	-0.209	6.90
Gold	11822	9.45	20.48	0.361	16.21
Crude	7315	10.44	39.60	-0.221	13.55
Comm	11089	2.64	12.70	-0.164	8.05

Table 2: Decriptive statistics of the data. Count is the number of non-missing daily observations. μ and σ are the annualized mean and standard deviations in percent, respectively. Skewness and Kurtotis are the skewness and kurtosis of the daily data.

	Log-likelihood Difference				Ranks			
	1	5	10	22	1	5	10	22
Agg. GARCH	–	-52.1	-74.7	-98.1	–	10.8	11.2	11.9
Agg. GARCH-RV	–	0.80	1.18	-0.67	–	5.24	5.52	6.88
Component GARCH	15.0	14.4	15.9	17.2	1.08	1.12	1.12	1.12
GARCH	0.54	2.25	4.81	6.58	5.88	3.92	3.52	2.48
HAR_{22}	-10.2	-14.8	-14.5	-15.0	9.92	9.72	9.88	9.92
HAR_{66}	6.57	3.74	4.29	3.55	3.52	3.84	4.32	3.96
MIDAS-β_{22}	-8.30	-12.9	-13.3	-14.3	8.04	7.84	8.32	8.44
MIDAS-β_{66}	6.21	4.23	4.84	3.61	3.82	3.50	3.46	4.06
MIDAS-exp_{22}	-10.2	-11.5	-3.97	-2.82	9.76	8.40	7.44	6.20
MIDAS-hyp_{22}	-8.90	-13.4	-13.7	-14.5	9.12	8.72	8.72	8.56
MIDAS-hyp_{66}	6.21	4.23	4.84	3.61	3.82	3.50	3.46	4.06
MA_{22}	-1154	-1348	-1376	-1411	12.6	13.0	13.2	13.6
MA_{66}	-320.0	-281.8	-276.2	-291.0	13.8	13.4	13.3	12.2
MA_{252}	-840.7	-542.3	-502.4	-492.7	12.5	11.9	11.5	11.6

Table 3: The left panel show the cross-series averge quasi likelihood difference relative to the median quasi log-likelihood for the series. The differences have been normalized to correspond to 10 years of daily data. The right panel reports the average rank for each model and horizon, where the ranks were computed series by series. Higher ranks correspond to larger quasi log-likelihoods.

	Agg.	Agg.	Comp.				MIDAS	MIDAS	MIDAS	MIDAS	MIDAS
	GARCH	GARCH$_{RV}$	GARCH	GARCH	HAR$_{22}$	HAR$_{66}$	β_{22}	β_{66}	exp	hyp$_{22}$	hyp$_{66}$
					Horizon 1						
Comp. GARCH	–	–	–	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.12
GARCH	–	–	0.52	–	0.04	0.24	0.08	0.56	0.00	0.04	0.40
HAR$_{22}$	–	–	0.76	0.28	–	0.64	0.04	0.72	0.16	0.16	0.56
HAR$_{66}$	–	–	0.32	0.08	0.00	–	0.00	0.04	0.00	0.00	0.12
MIDAS-β_{22}	–	–	0.76	0.32	0.04	0.52	–	0.72	0.16	0.16	0.60
MIDAS-β_{66}	–	–	0.16	0.08	0.00	0.00	0.00	–	0.00	0.00	0.12
MIDAS-exp	–	–	0.76	0.92	0.32	0.64	0.48	0.76	–	0.28	0.68
MIDAS-hyp$_{22}$	–	–	0.64	0.24	0.00	0.52	0.00	0.76	0.08	–	0.56
MIDAS-hyp$_{66}$	–	–	0.12	0.00	0.00	0.04	0.00	0.08	0.00	0.00	–
					Horizon 5						
Agg. GARCH	–	0.88	0.96	0.92	0.92	0.96	0.88	0.96	0.88	0.92	0.96
Agg. GARCH$_{RV}$	0.00	–	0.84	0.76	0.68	0.80	0.68	0.84	0.72	0.68	0.84
Comp. GARCH	0.00	0.00	–	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08
GARCH	0.00	0.00	0.40	–	0.04	0.24	0.04	0.56	0.00	0.04	0.32
HAR$_{22}$	0.00	0.00	0.52	0.32	–	0.44	0.12	0.60	0.12	0.32	0.44
HAR$_{66}$	0.00	0.00	0.24	0.00	0.00	–	0.00	0.12	0.00	0.00	0.20
MIDAS-β_{22}	0.00	0.04	0.52	0.20	0.04	0.36	–	0.68	0.12	0.24	0.40
MIDAS-β_{66}	0.00	0.00	0.04	0.04	0.00	0.00	0.00	–	0.00	0.00	0.12
MIDAS-exp	0.00	0.00	0.72	0.84	0.16	0.60	0.44	0.80	–	0.32	0.64
MIDAS-hyp$_{22}$	0.00	0.00	0.48	0.16	0.04	0.20	0.00	0.64	0.08	–	0.40
MIDAS-hyp$_{66}$	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.04	0.00	0.00	–
					Horizon 10						
Agg. GARCH	–	0.92	1.00	0.96	0.96	1.00	0.92	1.00	0.92	0.96	1.00
Agg. GARCH$_{RV}$	0.00	–	0.96	0.84	0.72	0.92	0.72	0.92	0.80	0.76	0.96
Comp. GARCH	0.00	0.00	–	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08
GARCH	0.00	0.00	0.40	–	0.00	0.32	0.00	0.56	0.00	0.00	0.36
HAR$_{22}$	0.00	0.00	0.56	0.20	–	0.52	0.00	0.68	0.08	0.32	0.48
HAR$_{66}$	0.00	0.00	0.20	0.00	0.00	–	0.00	0.12	0.00	0.00	0.16
MIDAS-β_{22}	0.00	0.00	0.56	0.20	0.12	0.48	–	0.64	0.08	0.32	0.48
MIDAS-β_{66}	0.00	0.00	0.00	0.00	0.00	0.04	0.00	–	0.00	0.00	0.16
MIDAS-exp	0.00	0.00	0.60	0.80	0.28	0.60	0.36	0.76	–	0.48	0.60
MIDAS-hyp$_{22}$	0.00	0.00	0.52	0.20	0.04	0.24	0.04	0.52	0.08	–	0.40
MIDAS-hyp$_{66}$	0.00	0.00	0.08	0.00	0.00	0.00	0.00	0.04	0.00	0.00	–
					Horizon 22						
Agg. GARCH	–	0.92	1.00	0.96	0.96	0.96	0.96	1.00	0.96	0.96	1.00
Agg. GARCH$_{RV}$	0.00	–	0.96	0.92	0.88	0.96	0.80	0.92	0.88	0.88	1.00
Comp. GARCH	0.00	0.00	–	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.12
GARCH	0.00	0.00	0.44	–	0.04	0.44	0.00	0.60	0.00	0.08	0.52
HAR$_{22}$	0.00	0.00	0.56	0.24	–	0.52	0.04	0.52	0.12	0.20	0.56
HAR$_{66}$	0.00	0.00	0.12	0.00	0.00	–	0.00	0.04	0.00	0.00	0.20
MIDAS-β_{22}	0.00	0.00	0.60	0.36	0.68	0.64	–	0.68	0.16	0.68	0.64
MIDAS-β_{66}	0.00	0.00	0.16	0.00	0.00	0.04	0.00	–	0.00	0.00	0.20
MIDAS-exp	0.00	0.00	0.64	0.80	0.56	0.72	0.40	0.68	–	0.48	0.64
MIDAS-hyp$_{22}$	0.00	0.00	0.56	0.24	0.16	0.48	0.04	0.48	0.16	–	0.52
MIDAS-hyp$_{66}$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	–

Table 4: Pairwise DMGW tests of the iterative forecasting methods for horizons 1, 5, 10 and 22. Table values indicate the percentage of times that the model in the column outperformed the model in the row.

	Losses				Ranks			
	1	5	10	22	1	5	10	22
Agg. GARCH	–	8	12	8	–	0	0	0
Agg. GARCH$_{RV}$	–	24	16	12	–	8	8	4
Comp. GARCH	92	96	96	100	4	76	76	76
GARCH	48	64	68	56	0	24	16	8
HAR$_{22}$	28	48	52	56	0	0	0	0
HAR$_{66}$	76	80	84	92	32	96	92	92
MIDAS-β_{22}	24	60	60	28	0	0	0	0
MIDAS-β_{66}	80	100	92	92	0	32	44	16
MIDAS-exp	16	20	36	28	0	0	4	4
MIDAS-hyp$_{22}$	24	60	64	52	0	0	0	0
MIDAS-hyp$_{66}$	88	100	96	96	4	72	80	72
MA$_{22}$	0	0	4	0	100	68	20	4
MA$_{66}$	0	0	0	4	0	0	0	0
MA$_{252}$	4	8	12	24	12	24	16	16

Table 5: Percentage of models that enter a Model Confidence Set with a family wise error rate of 10% for horizons 1, 5, 10 and 22 days. The left panel applies the MCS directly to losses, and the right panel applies the MCS to than rank of the losses.

	Avg. Losses				Avg. Losses (GLS)				Avg. Losses (Ranks)			
	1	5	10	22	1	5	10	22	1	5	10	22
Agg. GARCH	0	0	0	0	0	0	0	0	0	0	0	0
Agg. GARCH$_{RV}$	0	50	0	0	0	25	0	0	0	0	0	0
Comp. GARCH	100	100	100	100	100	100	100	100	0	25	50	75
GARCH	50	50	25	25	75	75	75	75	0	0	0	0
HAR$_{22}$	0	0	25	0	25	25	25	50	0	0	0	0
HAR$_{66}$	50	100	100	100	75	75	100	75	0	75	100	100
MIDAS-β_{22}	0	0	0	0	0	25	25	25	0	0	0	0
MIDAS-β_{66}	100	100	100	100	100	100	100	100	0	0	25	0
MIDAS-exp	25	0	25	0	25	25	25	25	0	0	0	0
MIDAS-hyp$_{22}$	25	25	50	25	50	25	25	25	0	0	0	0
MIDAS-hyp$_{66}$	100	100	100	100	100	100	100	100	0	25	50	50
MA$_{22}$	50	25	0	0	50	50	25	0	100	75	50	25
MA$_{66}$	0	0	25	50	0	0	0	50	0	0	0	0
MA$_{252}$	50	50	50	50	50	50	50	50	0	50	50	50

Table 6: Percentage of models that enter a Model Confidence Set for the losses grouped by category. The left panel uses the average loss when constructing the MCS, the middle panel uses a GLS-weighted average when constructing the MCS and the right panel uses the average rank when constructing the MCS. All report inclusion using a FWER of 10%.

	Avg. Losses				Avg. Losses (GLS)				Avg. Losses (Ranks)			
	1	5	10	22	1	5	10	22	1	5	10	22
Agg. GARCH	–	0.0	0.0	0.0	–	0.0	0.0	0.0	–	0.0	0.0	0.0
Agg. GARCH$_{RV}$	–	0.0	0.0	0.0	–	0.0	0.0	0.0	–	0.0	0.0	0.0
Comp. GARCH	100.0	100.0	88.3	48.0	100.0	100.0	100.0	90.7	0.0	91.2	89.9	54.3
GARCH	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0
HAR$_{22}$	0.0	0.1	0.0	2.2	0.0	0.0	0.1	0.4	0.0	0.0	0.0	0.0
HAR$_{66}$	0.9	10.9	15.6	10.5	2.3	9.8	18.4	8.4	0.0	91.2	70.5	100.0
MIDAS-β_{22}	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
MIDAS-β_{66}	14.8	96.2	88.3	25.6	23.4	76.3	91.7	40.3	0.0	0.0	0.1	0.0
MIDAS-exp	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MIDAS-hyp$_{22}$	0.0	1.8	3.7	4.1	0.1	0.1	0.1	0.4	0.0	0.0	0.0	0.0
MIDAS-hyp$_{66}$	19.3	96.2	100.0	100.0	23.4	76.3	91.7	100.0	0.0	88.3	100.0	54.3
MA$_{22}$	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	100.0	100.0	2.4	0.0
MA$_{66}$	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0
MA$_{252}$	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.4	0.0	3.3	2.4	0.0

Table 7: MCS p-values for incluđion when averaging losses across all models. The left panel contains results from a simple average of the losses and the right panel contains results for a GLS-type average of the losses.

					Estimation method: Returns				
	Avg. Losses			Avg. Losses (GLS)			Avg. Losses (Ranks)		
	5	10	22	5	10	22	5	10	22
Agg. GARCH	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Agg. GARCH$_{RV}$	0.4	2.1	98.0	0.6	7.0	64.6	0.0	0.0	0.1
Comp. GARCH	100.0	100.0	98.0	100.0	100.0	64.6	44.7	44.7	0.4
GARCH	0.4	7.8	98.0	0.8	15.0	100.0	0.0	0.0	0.0
HAR$_{22}$	6.6	35.6	54.1	4.3	7.0	3.1	0.0	0.0	0.0
HAR$_{66}$	32.8	47.1	100.0	43.2	60.6	89.5	100.0	100.0	100.0
MIDAS-β_{22}	0.4	0.5	1.6	0.0	0.2	1.1	0.0	0.0	0.0
MIDAS-β_{66}	1.4	0.6	0.4	3.2	3.8	0.0	0.0	0.0	0.0
MIDAS-exp	0.0	0.3	24.3	0.0	0.1	3.3	0.0	0.0	0.1
MIDAS-hyp$_{22}$	8.6	35.6	17.1	4.6	4.8	1.1	0.0	0.0	0.0
MIDAS-hyp$_{66}$	32.8	47.1	4.7	43.2	60.2	0.0	0.0	0.0	0.0

					Estimation method: Realized Variance				
	5	10	22	5	10	22	5	10	22
Agg. GARCH	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Agg. GARCH$_{RV}$	0.2	0.2	1.5	0.1	0.3	0.1	0.0	0.0	0.0
Comp. GARCH	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
GARCH	9.0	22.0	6.7	11.1	16.7	4.9	0.0	0.0	3.0
HAR$_{22}$	0.5	0.6	1.9	0.1	0.0	0.1	0.0	0.0	0.0
HAR$_{66}$	2.7	3.2	20.7	1.0	6.5	15.6	37.2	14.4	44.5
MIDAS-β_{22}	3.0	2.4	3.0	0.7	0.0	2.2	0.0	0.0	0.0
MIDAS-β_{66}	41.4	26.4	48.4	24.6	16.7	18.6	0.0	0.0	0.0
MIDAS-exp	0.0	0.0	0.3	0.0	0.0	0.1	0.0	0.0	0.0
MIDAS-hyp$_{22}$	9.0	22.0	48.4	2.3	6.5	15.3	0.0	0.0	0.0
MIDAS-hyp$_{66}$	41.4	34.4	48.4	24.6	16.7	18.6	2.3	3.8	0.0

Table 8: MCS p-values for includion when averaging losses across all models when forecasting using direct estimation using returns (top panel) or realized variance (bottom panel). The left panel contains results from a simple average of the losses and the right panel contains results for a GLS-type average of the losses.

	Iterative better			Direct better		
	Estimation method: Returns					
	5	10	22	5	10	22
Comp. GARCH	84.0	92.0	100.0	0.0	0.0	0.0
GARCH	84.0	80.0	92.0	0.0	0.0	0.0
HAR$_{22}$	76.0	92.0	92.0	0.0	0.0	0.0
HAR$_{66}$	84.0	92.0	96.0	0.0	0.0	0.0
MIDAS-β_{22}	76.0	92.0	96.0	0.0	0.0	0.0
MIDAS-β_{66}	88.0	100.0	96.0	0.0	0.0	0.0
MIDAS-exp	80.0	88.0	92.0	0.0	0.0	0.0
MIDAS-hyp$_{22}$	76.0	92.0	92.0	0.0	0.0	0.0
MIDAS-hyp$_{66}$	84.0	96.0	96.0	0.0	0.0	0.0
	Estimation method: Realized Variance					
	5	10	22	5	10	22
Comp. GARCH	8.0	16.0	36.0	0.0	0.0	0.0
GARCH	0.0	4.0	4.0	0.0	16.0	16.0
HAR$_{22}$	12.0	12.0	8.0	0.0	4.0	24.0
HAR$_{66}$	20.0	24.0	44.0	4.0	0.0	0.0
MIDAS-β_{22}	24.0	24.0	4.0	12.0	12.0	36.0
MIDAS-β_{66}	36.0	28.0	32.0	0.0	4.0	8.0
MIDAS-exp	0.0	0.0	0.0	4.0	16.0	24.0
MIDAS-hyp$_{22}$	24.0	24.0	8.0	4.0	8.0	32.0
MIDAS-hyp$_{66}$	24.0	28.0	48.0	4.0	0.0	0.0

Table 9: Percentage of DMGW test rejecting in favor of iterative version (left columns) or direct foreasts (right columns) . The top panel uses direct forecasts estiamtes using $h-$period returns. The bottom panel uses direct forecasts where the parameters were estimated on the h-day realized variance.

	Iterative better			Direct better		
	5-year Estimation Window					
	5	10	22	5	10	22
Comp. GARCH	8.0	36.0	60.0	4.0	0.0	0.0
GARCH	4.0	28.0	24.0	0.0	0.0	0.0
HAR_{22}	20.0	24.0	12.0	0.0	12.0	8.0
HAR_{66}	44.0	40.0	68.0	0.0	0.0	0.0
MIDAS-β_{22}	24.0	20.0	4.0	0.0	4.0	12.0
MIDAS-β_{66}	36.0	44.0	48.0	0.0	0.0	0.0
MIDAS-exp	16.0	24.0	12.0	0.0	4.0	4.0
MIDAS-hyp_{22}	24.0	32.0	32.0	4.0	4.0	12.0
MIDAS-hyp_{66}	20.0	40.0	64.0	0.0	0.0	0.0
	20-year Estimation Window					
	5	10	22	5	10	22
Comp. GARCH	4.0	4.0	4.0	12.0	16.0	16.0
GARCH	0.0	0.0	0.0	24.0	16.0	28.0
HAR_{22}	4.0	4.0	4.0	16.0	36.0	72.0
HAR_{66}	16.0	8.0	16.0	0.0	0.0	12.0
MIDAS-β_{22}	8.0	8.0	4.0	16.0	32.0	84.0
MIDAS-β_{66}	20.0	20.0	8.0	4.0	12.0	24.0
MIDAS-exp	0.0	0.0	0.0	24.0	24.0	40.0
MIDAS-hyp_{22}	4.0	4.0	4.0	24.0	44.0	84.0
MIDAS-hyp_{66}	16.0	12.0	24.0	8.0	0.0	4.0

Table 10: Percentage of DMGW test statistics that prefer the iterative forecast (left panel) or the direct forecast (right panel) when the estimation window is 5 years of daily data (top panels) or 20 years of daily data (bottom panels). The parameters used to produce the direct forecasts were estimated using the realized variance-based quasi likelihood.

	Losses			Ranks		
	5	10	22	5	10	22
	Direct (RV)					
Comp. GARCH	96	96	80	72	84	60
HAR_{66}	60	76	60	68	60	40
MIDAS-β_{66}	76	84	72	16	24	16
MIDAS-hyp_{66}	84	84	72	40	52	24
	Iterative					
Comp. GARCH	96	96	100	76	80	80
HAR_{66}	84	88	92	84	84	84
MIDAS-β_{66}	100	100	92	32	48	12
MIDAS-hyp_{66}	92	92	96	76	84	80

Table 11: Percentage of models that enter a Model Confidence Set with a family wise error rate of 10% for horizons 1, 5, 10 and 22 days. The left panel applies the MCS directly to losses, and the right panel applies the MCS to than rank of the losses. The top set of models were direct forecasts using parameters estimated with realized variance and the bottom used iterative forecasts.

B Figures

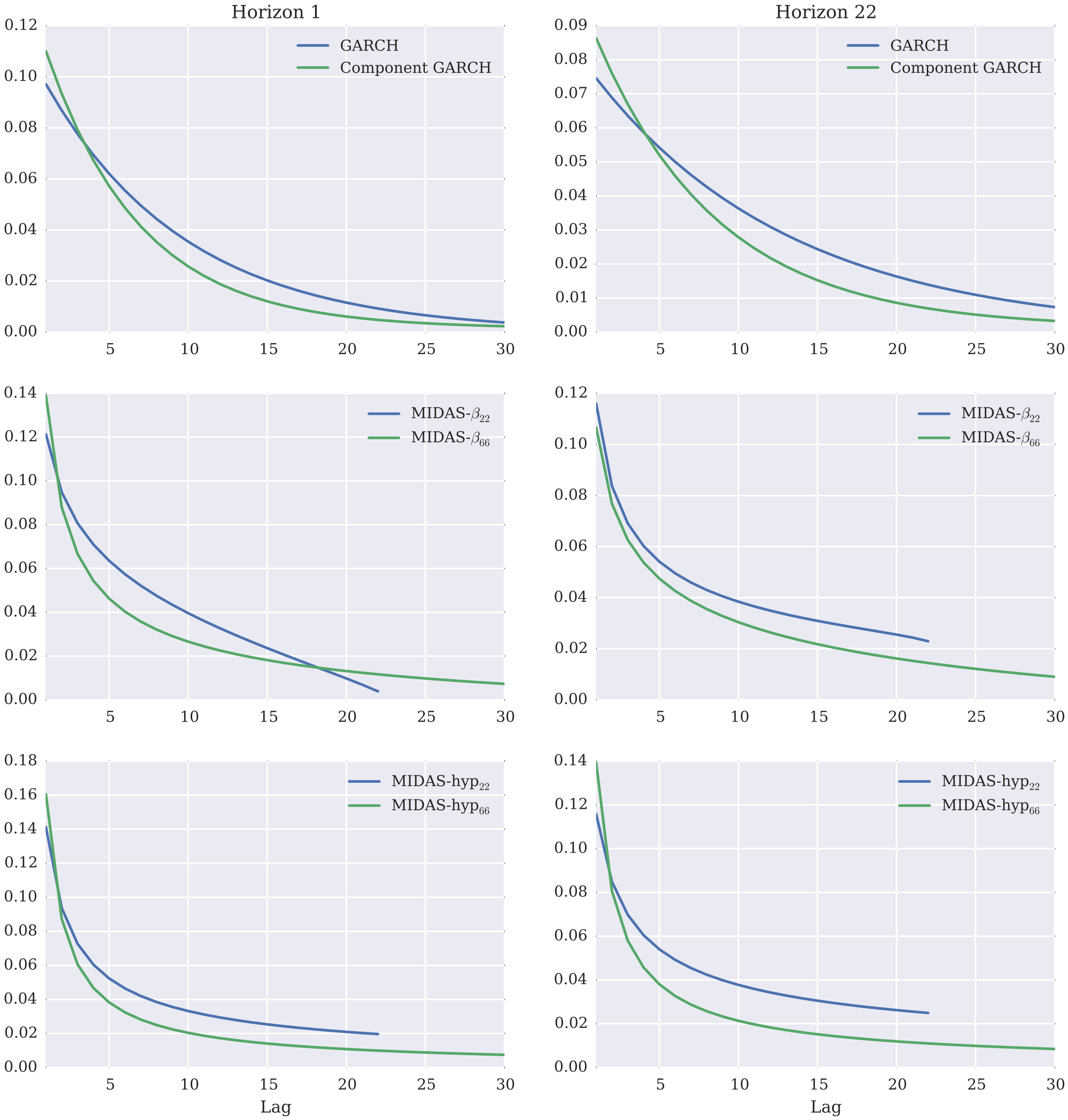

Figure 1: Weight given to observation at lag j for the first 30 lags. The left panel show the weights when the model is optimized to predict on one-day ahead volatility and the right panel shows the weights when optimized to predice 22-day ahead volatility. All weights based on full sample estimates of the VWM series.