A reality check on the GARCH-MIDAS volatility models

Nader Virk[¬], Farrukh Javed[*], Basel Awartani[~] and Stuart Hyde[^]

Abstract

We employ a battery of model evaluation tests for a broad set of GARCH-MIDAS models and account for data snooping bias. We document that inferences based on standard tests for GM variance components can be misleading. Our data mining free results show that the gain of macro-variables in forecasting total (long-run) variance by GM models is overstated (understated). Estimation of different components of volatility is crucial for designing differentiated investing strategies, risk management plans and pricing derivative securities. Therefore, researchers and practitioners should be wary of data-mining bias, which may contaminate a forecast that may appear statistically validated using robust evaluation tests.

Keywords: Forecasting, GARCH-MIDAS models, component variance forecasts, macro-variables, data snooping.
JEL classification: C32, C52, G11, G17.

¬ Huddersfield Business School, Huddersfield, UK. Email: N.S.Virk@hud.ac.uk

* Corresponding author. Lund University, Sweden. Email: farrukh.javed@stat.lu.se

~ King Fahd University of Petroleum and Minerals, KSA. Email: basel.awartani@kfupm.edu.sa

^ The University of Manchester Alliance Manchester Business, UK. Email: stuart.hyde@manchester.ac.uk

1 Introduction

The new class of two-component volatility models pioneered by Engle, Ghysels and Sohn (2013) incorporates a long-run variance component in volatility modelling.[1] The novel GARCH-MIDAS (GM) models combine transitory/short-term GARCH volatility with the variability of the low-frequency information channels, predominantly economic sources. This model is suitable for modelling short and long-term volatilities in financial markets.[2]

Prior literature has employed a range of aggregate variables to assess their linkages to stock market volatility. The prominent aggregate variables are industrial output (Officer, 1973), interest ratess (Shiller 1981a), industrial production and inflationary changes (Schwert 1989), and monetary policy variables, including money supplies and exchange rate fluctuations (Eichengreen and Tong 2003), oil price shocks (Degiannakis, Filis, and Kizys 2014) and unemployment changes (Asgharian et al. 2013). With the exception of Asgharian et al. (2013) these studies investigated equity volatility at mismatched time frequency to low frequency aggregate variations. Noticeably, one shortcoming of the weak linkages between equity volatility and macroeconomic variables has been lack of appropriate methodological approaches. Engle et al. (2013) solve this issue by proposing the GM two-component model and providing robust evidence on the importance of macroeconomic data to forecast the US equity volatility.

Asgharian et al. (2013) and Conrad and Loch (2015) provide additional tests for volatility forecasting using GM models while expanding on the set of economic variables originally employed in Engle et al. (2013). The evidence in both studies reinforces the results in seminal work on GM models and documents that GM models using principal component (PC) analysis can enhance the forecasts of the GM for both the total and the long-term variance. Conrad and Loch (2015) report that the term spread, housing starts, and consumer sentiment are the leading predictive variables for US stock volatility.[3] In forecasting long-run US equity volatility, Lindblad (2017) shows that incorporating sentiment adds to total and long-term variance predictability, whereas Conrad and Kleen (2020) reinforce the fact that housing starts are an excellent predictor

[1] The multiplicative component structure of these models combines the GARCH volatility with the long-run approximations for long-run variance – estimated by the innovative Mixed Data Sampling (MIDAS) approach of Ghysels, Santa-Clara and Valkanov (2004, 2006). Given the importance of long-memory dependence in financial market volatility, the long-run variance smoothing by the MIDAS regressions – forecasting mismatched time frequency (e.g. monthly) variance from high frequency data points such as past daily squared returns or economic variables – has opened up new frontiers in examining the role of long memory processes in shaping financial volatility and how cross-market long-run variances and correlations depict inter- and intra-market integration patterns.

[2] We invariably refer long-term variance to trend/secular/MIDAS component or variance to imply the same throughout the paper.

[3] In addition to modelling conditional variance, the extensions that build on the GM framework have been utilised in studying the relationships between oil and stock market volatilities, oil-stock correlation, stock-bond correlation, oil-macroeconomic relationships, European equity market integration patterns and the effect of investor sentiment on US stock-bond correlation patterns (Conrad, Loch and Rittler 2014; Asgharian, Christiansen and Hou 2016, Pan, Wang, Wu, and Libo 2017, and Virk and Javed 2017, Fang, Yu and Huang 2018).

for stock volatility. Pan et al. (2020) report the importance of macroeconomic activities on the aggregate volatility market through a newly proposed model involving jumps within the GM sets. In sum, the evidence suggests clear advantages to using macroeconomic information in predicting volatility in the context of GM models.

In the literature, many researchers use GM models in volatility prediction. Their work has been motivated by the lack of appropriate models that combine the short and the long components of the latent volatility process.[4] However, three issues/gaps need scrutiny when examining the overall gains of modelling volatility within the class of GM models. First, most GM volatility modelling evidence uses US equity data. There has been little or no evidence of GM models for volatility modelling from other global markets. Second, the emerging evidence displaying the utility of GM models compares the forecasts of GM with macro information with the predictions of a GM model that is a function of lagged monthly realised variance (RV).[5] Although unbiased and consistent, RV is still a noisy estimate of ex-post volatility. Andersen and Bollerslev (1998) illustrate that lower frequency forecasts of the latent volatility may benefit from constructing factors at higher frequencies to mitigate the idiosyncratic noisiness of the RV measure.[6] Conjecturing from their evidence, we hypothesise that compuring cumulative monthly RV through daily squared returns will provide the same benefit in smoothing variance in our work.

Thus, merging our conjecture with the findings dislplaying a weak relationship between stock volatility and aggregate variables (Officer 1973, Shiller 1981b, Christie 1982, Schwert 1989, among others), leads us to question the evolving evidence on GM models. Are there actual GM specifications performing better than the standard GM benchmark?[7] More succinctly, we inquire if (i) the evidence from GM models with macro variables differs from that reported in the existing literature and (ii) forecasts that incorporate macro information outperform forecasts of the GM model that conditions on lagged RV to approximate latent volatility.

We identify that forecasting exercises in studies using GM models typically use weak yet informative tests to draw inferences about the relative performance of the GM variance forecasts

[4] There are numerous stylized facts documented about financial market volatility, such as clustering, leverage effect, mean reversion and co-movement of volatilities among assets and across markets (Cont 2001).

[5] Potentially, daily RV can be a noisy estimate for forecasting volatility at high frequencies (Andersen and Bollerselv 1997, 1998). However, we argue that the same is not applicable to the low frequency RV estimates when we model low frequency variance components at monthly, quarterly and bi-annual along with the total variance forecasts.

[6] To this extent, Andersen and Bollerselev (1998) show that a well-specified, GARCH-type volatility filter smooth unconditional realised variance/volatility (RV), from high frequency intraday data, produces precise inter-daily predictions i.e., a latent volatility factor from high frequency intraday data improves out-of-sample forecasts at inter-daily variance predictions.

[7] Although the importance of economic sources influencing market volatility is undeniable, there are limitations in the modelling the contribution of economic variables to the total and long-run variance forecasts. These include identification of the aggregate variable(s) that contributes to the evolution of financial market variance and prediction of financial market volatility on the basis of an information set that is prone to measurement errors and revisions compared with other variables to proxy long-run variations in the conditional variance such as RV.

against the benchmark models.[8] These tests mostly use the mean squared error (MSE) type loss function. In the spirit of Engle et al. (2013), a relative MSE ratios test identifies the efficiency gains of the competing model against the benchmark model or vice versa. However, Asgharian et al. (2013) have assessed the forecasting merits of the GM model using the Diebold and Mariano (DM, 1995) test – a robust statistical test for forecast evaluation.[9] We also note that prior studies, except Asgharian et al. (2013), have usually generalised the evidence from GM models for total volatility to long-run variance forecasts.

Finally, we observe that given the variety of economic sources as well as financial information variables, there are several candidate GM models to forecast volatility. The repetitive use of the same equity data is bound to shadow the reliability and accuracy of the GM forecasting volatility inferences. Therefore, in consideration to data mining issue, we are sceptical about the reliability of prior evidence on the forecasting superiority of GM models and for not differentiating between total and long-run variance forecasts. To provide rigour to these forecasting comparisons using the GM models, we argue that model comparisons should account for statistical tests that control for data snooping biases before inferring the value of GM macro models in volatility prediction.

These identified research gaps motivate our work. To limit our study, we note that Engle et al. (2013) specify two types of GM models. The first type is the one-sided filter, where long-term volatility of daily stock returns is expressed as a weighted average of lagged values of lower-frequency financial/macroeconomic variables. The other model type is two-sided filters that use lagged and future values of the MIDAS input variables. Our work only studies the first type of GM model. We provide new evidence in two ways. First, we replicate research using the one-sided GM models across four leading global equity markets, i.e., France, Germany, the UK and the US. Second, we assess the robustness of earlier evidence using tests that account for data snooping bias.

To do this, we investigate the gains of several macro-variables in the GM models to forecast total and long-run variance components using a common GM benchmark that only uses monthly RV. We carry out informative and robust statistical tests that correspond with prior research on GM models. Furthermore, we use a far more extensive set of macroeconomic variables than all previous studies that assessed the importance of information coming from different aggregated channels. Only Conrad and Loch (2015) comes close in this respect. Nonetheless, our work

[8] This applies to tests for both total variance and long-run variance components regardless of whether the benchmark model is the GM model that smooths realised variance in MIDAS regressions (Engle et al. 2013 and Conrad and Loch 2015) or the baseline GARCH (1, 1) specification (Asgharian et al. 2013 and Conrad et al. 2015).

[9] The pairwise DM test examines the equal predictability (EPA) of the alternate model against the benchmark.

extends evidence on GARCH-MIDAS modelling and forecasting for a cross-section of four developed equity markets and thus provides out-of-sample evidence.

Second, we employ powerful (multiple/joint) tests that account for the data mining issues i.e., the White (2000) reality check (RC) test and the Hansen (2005) superior predictive ability (SPA) test. Both tests assess the SPA of the benchmark model when contrasted against a host of alternative models and are robust against data mining biases. Specifically, these tests evaluate the increase in the probability of finding SPA among the competing models when the number of competing models increases (White, 2000 and Hansen, 2005).[10] Our evidence accounting for the data snooping bias for the class of GM models with one benchmark is the first in this regard. Here it is important to note that contemporary work by Lindblad (2017) and Conrad and Kleen (2020) have made use of the model confidence set (MCS) of Hansen et al. (2011) for forecast evaluation of GM models.[11] The evidence in the latter study shows that the GM RV model is outperformed by models that use macroeconomic information. Furthermore, the lag structure for the input MIDAS variables is given by the weighting scheme – unrestricted or restricted – adopted for the beta polynomial by which MIDAS variance evolves. Thus, the estimation of long-run variance depends on the weights given to the input variables' lagged values in the MIDAS filter. Conrad and Loch (2015) confirm the evidence in Engle et al. (2013) that the optimal weighting scheme for the GM-RV (benchmark) model is the restricted one. The former study does show that some macroeconomic variables require an unrestricted scheme while RV does not.

Nonetheless, there are studies that adopt an ad-hoc restricted weighted scheme even for macroeconomic variables in the MIDAS filter, such as Asgharian et al. (2013) and Conrad et al. (2014), among others. We expect that imposing this weighting scheme for other low-frequency input variables may not optimally forecast long-run variance given the flexible weight convergence, as shown in Conrad and Loch (2015). To add value to our empirical analysis, we estimate all GM models using unrestricted and restricted weighting schemes for the MIDAS variance component across cross-sections of four developed equity markets – just not the US equity volatility.

In summary, we carry out a large-scale empirical exercise for the class of GM models, using one-sided filters across four large global equity markets. Through this exercise, we investigate the

[10] Here we note that the model comparisons in our study are for long-run variance and total variance as provided by the GM models. The comparability tests for volatility predictive ability tests for short run variance of GARCH type models have already been studied extensively, see Lunde and Hanen (2005) and Gonzalez, Lee and Mishra (2004) and others, and therefore we refrain from reporting that evidence to conserve space.

[11] As mentioned earlier, initial work on GM modelling has used various pairwise tests e.g., Diebold-Mariano tests. Recent work on GM modelling and forecasting is transcending to evaluate its gains using stronger statistical tests. For example, Lindblad (2017) relies on the model confidence set approach of Hansen et al. (2011). Similarly, Conrad and Kleen (2020) also use MCS test for GM models. However, both the studies use only the US data.

gains and efficacy of volatility forecasts when total and long-run volatility evolves, using macro information in the MIDAS filter relative to the benchmark model while controlling for data snooping biases. We follow Engle et al. (2013) and adopt the GM model that uses rolling window (RW) monthly RV in the MIDAS filter as the benchmark model, referred to as GM-RV model hereafter.

Our analysis covers daily equity and monthly macro data from 1999 to 2022. We estimate 27 GM models for every market, resulting in the same number of forecasts for the total variance and trend component. This number is doubled since we use two weighting schemes in the MIDAS regressions. Using the informative MSE ratio test of Engle et al. (2013) and the Diebold and Mariano (1995) test, our results are congruent with the findings from the US equity data. All comparisons use one-step ahead forecasts i.e., pseudo-out-of-sample (POS) GM variance forecasts (unless otherwise stated).

Our results show that the evidence using the MSE ratio and DM test overstates the gains of the total variance forecasts coming from GM macro models when contrasted against the benchmark model forecast. Several competing models outperform the benchmark model's MSE. The DM test concurs when it uses an unrestricted weighting scheme. However, with the restricted weighting gains are limited to forecasts for the MIDAS/long-run variance component only – an aspect that is typically overlooked in the prior GM model evaluation literature. These summary results from the DM test essentially hold across all markets as we account for data snooping problems. The only exception is the fact that joint tests show that no competing model has SPA over the benchmark model total variance forecasts regardless of what weighting scheme is adopted.

We examine the consistency of our results by using alternate variance proxies and an out-of-sample forecasting scheme.[12] The generality of our results for the POS is maintained using alternate variance proxies. The OS rolling forecasting comparisons results – using any type of variance proxy – confirm our POS evidence. There are exceptions, however. We note that with the OS forecasting procedures, the total (long-run) variance forecast evaluations are sensitive to the type of weighting scheme adopted, which endorses our scepticism on model comparison exercises using GM models with restricted weights. Our evidence shows that results from a particular weighting scheme cannot be generalised for forecasts coming from GM models with other weighting structures, which are in accordance with Conrad and Loch (2015), where the authors estimated all models with a restricted and unrestricted weighting scheme and used the

[12] The POS forecasting comparisons are one step ahead only and can be taken as equivalent to a fixed forecasting scheme, while our OS joint tests give multiple steps ahead forecasts. More factually one year at time, for details see section 5.4.

likelihood ratio tests to identify the appropriate specification. This also applies to using alternative forecasting schemes such as one step ahead or multiple steps ahead forecasts. Most importantly, we should scrutinise the usefulness of the GM concerning its two distinctive forecasts, i.e., total variance and long-run variance – neither can substitute for the other. Finally, the use of powerful tests is suggested to examine the forecasting gains of a particular model over the benchmark model, e.g., when p-values of robust statistical tests reject the superiority of the forecasting performance of the benchmark model.

Overall, we conclude that, although macroeconomic information may not improve the accuracy of total GM volatility forecasts, there is overwhelming evidence of its usefulness in improving the long-term equity volatility forecasts. Nonetheless, guided by the sum of our results from data mining free POS and OS forecasting comparisons, we broadly find that the existing literature over(under)-states the gains for forecasting GM total (long-run) variance with aggregate variables. Our results have value to academics, investors and practitioners alike when we know that financial volatility has different components. The emphasis is on the econometrician to be rigorous and liberal in their search for forecasts for the other components of conditional financial volatility. If the GM-RV model has SPA over competing models it does not translate into superiority for the low-frequency variance component without limiting the evolution of the secular variance part. We know investors, money and risk managers with heterogeneous investment and risk management needs require volatility forecasts for different time horizons, investment mandates, geographical tilts and informational flow heterogeneity.

The rest of the paper is organised as follows. Section 2 discusses the GM methodology. Section 3 details the forecasting comparison testing procedures. Section 4 specifies data, sources, and summary statistics. The results are summarised in section 5. And section 6 concludes the findings and implications of the work.

2 GARCH-MIDAS models

In this section, we outline the GARCH-MIDAS methodology. The two-component GM volatility models break down the total variance for a financial asset into a short-term transitory component and trend/secular component. The multiplicative total variance is modelled by a unit variance GARCH (1, 1) and a secular component, which is estimated by the MIDAS filter.[13]

[13] The estimates for secular volatility can be obtained through several economic, financial and sentiment related variables.

To set up the model notations, we assume the compounded return for a price series on day i in month t is $r_{i,t} = \mu + \sqrt{\tau_t g_{i,t}}\varepsilon_{i,t}, \forall\, i = 1, \dots, N_t$ where $\varepsilon_{i,t}|\Phi_{i-1,t} \sim N(0,1)$ and $\Phi_{i-1,t}$ is the information set until day $i-1$ in period t (month in our case). The compounded return series have mean μ as its location parameter and its scale parameter, i.e. the total conditional variance $\sigma_t^2 = \tau_t g_{i,t}$ comprises a transitory (GARCH) component $g_{i,t}$ and a long-run (MIDAS) component τ_t.

Engle et al. (2013) specify the short-term volatility component as a GARCH (1, 1) process:

$$g_{i,t} = (1 - \alpha - \beta) + \alpha \frac{(r_{i-1,t} - \mu)^2}{\tau_t} + \beta g_{i-1,t} \qquad (1),$$

where $\alpha > 0, \beta \geq 0$ and $\alpha + \beta < 1$.

Here, the long-term volatility process, τ_t, can evolve by a range of low frequency variables such as macroeconomic variables. Under the MIDAS setting, this component is estimated:

$$\tau_t = \theta_0 + \theta_1 \sum_{k=1}^{K} \phi_k\,(w_1, w_2)\; X_{t-k}\,, \qquad (2)$$

where X is any variable of interest, however, just as in any another regression, it can accommodate more than one variable as well, around which the secular component should be determined. Following Conrad and Loch (2015) and Engle et al. (2013), the volatility forecast for a specific day i in month t is

$$E\left[r_{i,t} - E(r_{i,t}|\,\Phi_{N^{(t-1)},t-1})\right]^2 = E\left[g_{i,t}\tau_t\,\varepsilon_{i,t}^2|\,\Phi_{N^{(t-1)},t-1})\right] = \tau_t E\left[g_{i,t}|\,\Phi_{N^{(t-1)},t-1})\right] = \tau_t,$$

assuming the $E\left(g_{i,t}|\,\Phi_{N^{(t-1)},t-1}\right)$ converges to unity, i.e., equal to its unconditional expectation, $E_{t-1}\left(g_{i,t}\right) = 1$, provided i is large.[14]

To estimate the long-term variance at t, $\phi_k\,(w)$ is a smoothing function that provides a weighting scheme for the lagged values of the variable(s) of interest in the MIDAS filter. We choose the beta smoothing function that determines optimal weights with which the lags of the input variables in the MIDAS regressions are going to shape the secular volatility through the parameter θ_1. It is specified:

$$\phi_k\,(w_1, w_2) = \frac{(k/K)^{w_1-1}(1-k/K)^{w_2-1}}{\sum_{j=1}^{Kv}(j/K)^{w_1-1}(1-j/K)^{w_2-1}} \qquad (3)$$

where w_1, w_2 are weights to be estimated.

[14] Note that the above convergence only holds for large i, more details can be found in Conrad and Loch (2015) and Engle et al. (2013).

The choice of beta smoothing function is motivated by the fact that it involves two weights w_1 and w_2,which gives more flexibility in accounting for different features in the data. It can be shown that the chosen weight function is monotonically decreasing as long as w_1 is equal to one. Given $w_1 = 1$ and increasing w_2 gives a larger weight to the most recent observations while a $w_1$larger than one gives a lower weight to the most recent observations. More discussion, in this regard, can be found in Asgharian et al. (2013).

Using the flexible/unrestricted beta smoothing function, the long-term volatility of daily returns in equation (2) is expressed as a weighted average of lower-frequency financial and/or macroeconomic variables. This beta-polynomial is independently estimated for each MIDAS regression and for each input variable therein.

Studies have used restricted version of the above weighting scheme by fixing $w_1 = 1$ (Engle et al. 2013, Asgharian et al. 2013, Conrad et al. 2014, among others): $\phi_k\ (w_2) = \frac{(1-{}^{k}/_{K})^{w_2-1}}{\sum_{j=1}^{K}(1-{}^{j}/_{K})^{w_2-1}}$.

Ghysels, Sinko and Valkanov (2007) report that the unrestricted smoothing scheme allows for a hump-shaped decaying pattern. For restricted weighting scheme, the fixed weight of $w_1 = 1$ ensures a decaying pattern whereas the size of w_2 determines the speed of decay: large (small) values of w_2 generate an accelerating (decelerating) decaying pattern for the lagged values of input variable(s) in the MIDAS filter. It important to note that the unrestricted case is flexible in providing/estimating large weights for distant lags of MIDAS input variable (i.e. the hump shape decline), fixing$w_1 = 1$ makes it a strictly declining case for the number of lags involved in the MIDAS smoothing.

To clarify, we note that the parameter space for the baseline GM model using RV as input variable results in $\Theta = \{\mu, \alpha, \beta, \theta_0, \theta_1, w_2\}$ when we use the restricted weighting scheme. It is understood that the parameter space for the GM-RV model with the unrestricted weighting scheme will result in $\Theta = \{\mu, \alpha, \beta, \theta_0, \theta_1, w_1, w_2\}$ i.e., one more parameter estimate than the restricted case that fixes$w_1 = 1$. The parameter space changes accordingly for GM specifications that have more than one exogenous variable to smooth secular component. For example, the unrestricted parameter space for two exogenous variables in the MIDAS regression will become $\Theta = \{\mu, \alpha, \beta, \theta_0, \theta_1, \theta_2, w_{1,\theta_1,}, w_{2,\theta_1,}, w_{1,\theta_2}, w_{2,\theta_2}\}$. Analogously, the restricted version will contain $2p$ less parameters, where p represents the number of input variables in the MIDAS regression. In the scope of our work, we estimate all the models using both weighting schemes. First, we let each MIDAS regression search for the properties of exogenous variable(s) in the MIDAS regressions

using the flexible weighting scheme i.e., we estimate w_1 and w_2. We then follow this by replicating all the MIDAS regressions using a constrained weighing scheme.

For clarity, we specify the baseline long-run component (τ_t), which uses rolling window (RW) monthly realised volatility, is available at daily frequency (we drop the 'i' subscript from the GM-RV equations for ease of expression):

$$\tau_t = \theta_0 + \theta_1 \sum_{k=1}^{K} \phi_k (w_1, w_2)\ RV_{t-k} \quad (4)$$

where the rolling window $RV_t = \sum_{i=1}^{N} r_{i,t}^2$, $N = 22$ approximates monthly realised volatility, and K lags of the input variable(s), are utilised to smooth trend component.[15] The relation in equation (4) can also be specified as

$$log\ (\tau_t) = \theta_0 + \theta_1 \sum_{k=1}^{K} \phi_k (w_1, w_2)\ RV_{t-k} \quad (5)$$

Following Engle et al. (2013), we adhere to the log-specifications for all the models: a log version of GM-RW RV or simply GM-RV is directly comparable to the competing GM specifications that involve macroeconomic variables. The log transformation of equation (4) guarantees the non-negativity of the conditional variances when the input variables (e.g. term structure of interest rates, industrial production) can take negative values. The transformation of $\log(\tau_t)$ specification gives τ_t which is an exponential estimate of the right side of the equation (see Engle et al. 2013 eq. 19 on page 781).

In a similar fashion, one can construct the long-term volatility component using the levels of macroeconomic variables: X_{t-k}^{l} denotes the level of X input variable in the MIDAS filter:

$$\log(\tau_t) = \theta_0 + \theta_1 \sum_{k=1}^{K} \phi_k (w_1, w_2)\ X_{t-k}^{l} \quad (6).$$

Or further extend this model by incorporating the variance of macroeconomic information as well:

$$\log(\tau_t) = \theta_0 + \theta_1 \sum_{k=1}^{K} \phi_k (w_{1\theta_1}, w_{2,\theta_1})\ X_{t-k}^{l} + \theta_2 \sum_{k=1}^{K} \phi_k (w_{1,\theta_2}, w_{2,\theta_2})\ X_{t-k}^{v} \quad (7)$$

Where, X_{t-k}^{v} is the variance of X input variable in the MIDAS filter. For generality, a specification by adding RV together with the macroeconomic information, both at level and variance is

$$\log(\tau_t) = \theta_0 + \theta_1 \sum_{k=1}^{K} \phi_k \left(w_{1\theta_1}, w_{2,\theta_1}\right)\ RV_{t-k} + \theta_2 \sum_{k=1}^{K} \phi_k (w_{1,\theta_2}, w_{2,\theta_2})\ X_{t-k}^{l} +$$
$$+ \theta_3 \sum_{k=1}^{K} \phi_k (w_{1\theta_3}, w_{2,\theta_3})\ X_{t-k}^{v} \quad (8)$$

Using the models implied by equations 5-8, in total 26 MIDAS and GM log specification models are iterated to compute model forecasts for comparison with the benchmark GM-RV model (a list is presented in Table 1). For simplicity and differentiation, we notate the benchmark model that

[15] In our baseline specification, we use RW-RV in the MIDAS smoothing at daily frequency: each daily realised variance is the rolling sum of 22-daily squared returns. Whereas the long-run variance smoothing for all other GM specifications including macro variables and/or principal components is at monthly frequencies only: long-run variance component changes at monthly frequency and stays constant for the days in a month.

involves RW-RV as GM-RV model and all remaining (competing M) models are referred by GM models.

3 Model forecast comparisons

In our work, the conditional total and long-run variance forecasts from the competing M GM models are compared against the GM-RV forecasts, respectively $\sigma^2_{t,GM-RV}$ and $\tau_{t,GM-RV}$. The parameters of the volatility models are estimated using the Q sample observations. These estimates are then used to make one step ahead pseudo out-of-sample forecasts, in our case: $P = 1$, which then are taken to different model comparisons tests.[16] The model comparison tests are undertaken for both the total variance forecasts at daily frequency and the long-run variance forecasts at monthly frequency. The forecast errors are calculated with respect to monthly RV, from daily stock return series for each market for the secular components and using squared returns for the total variances. So, when evaluating model forecasts, we proxy latent conditional variances by squared returns for daily data, and use the sum of daily squared returns in month t i.e. $\hat{\sigma}^2_t = RV$ for monthly data.

Consider that there are M competing models: $M = 1,2,\dots,m$ and GM-RV is the benchmark model. Each model m provides a forecast or series of forecasts $\{g_{m,t}\tau_{m,t}\}_{t=1}^{P}$which are compared to $\{\hat{\sigma}^2_t\}_{t=1}^{P}$, yielding a MSE based loss function L. Each model leads to a sequence of losses/forecast errors i.e. the losses for benchmark model are defined as, $L_{GM-RV,t} \equiv L(\hat{\sigma}^2_t, g_{GM-RV,t}\tau_{GM-RV,t})$ and losses for a competing k model are defined as, $L_{m,t} \equiv L(\hat{\sigma}^2_t, g_{m,t}\tau_{m,t})$. Using these quadratic losses, we conduct numerous model comparisons tests that are briefly explained below.

3.1 MSE ratios

This informative forecasting comparison metric is presented in Engle et al. (2013) to assess model performance of the GARCH-MIDAS models. The statistic, which is the ratio of the MSE of conditional variance forecasts of the m-competing model e.g. $RV^{\sigma^2}_{m,MSE}$ relative to MSE of the total variance forecast of the benchmark model i.e. GM-RV model, is defined as,

$MSE\ ratio \equiv RV_{m,MSE}/RV_{GM-RV,MSE}$, for $M = 1,2,\dots,m$.

These ratios are computed for both long-term and the total variances of all the competing models with reference to corresponding MSE of the benchmark model. Given that the benchmark model is in the denominator, a ratio of less than one implies an improvement over the benchmark model.

[16] The POS, which could also be regarded as in-sample fitting, is a terminology that follows from Engle et al. (2013) and is adopted for ease of comparison.

In other words, a ratio lower than 1 indicates a lower loss of accuracy in the sample of competing forecasts compared to the GM-RV benchmark model.

3.2 Pairwise tests of equal predictive ability

Tests for equal predictive ability (EPA), in a general setting, were proposed by Diebold and Mariano (1995) and West (1996). The DM test is used to compare the prediction accuracy of two competing models with the null hypothesis of no difference between the accuracy of two competing forecasts.

The relative performance (RP) measure between the loss function, which in our case is quadratic, of two competing models is:

$$RP_{k,t} \equiv L_{m,t} - L_{GM-RV,t}, \text{ for } M = 1,2, \dots, m \quad (9)$$

While the test proposed by Diebold and Mariano (1995) focuses on EPA and is pairwise, testing whether a particular forecasting procedure is superior to the alternative forecasts requires a test of superior predictive ability (SPA). Therefore, we carry out robust tests of superior predictive ability proposed by White (2000) (hereafter, RC test) and Hansen (2005) (hereafter, SPA test). These tests are explained below.

3.3 Tests for superior predictive ability

Utilising the terminologies set earlier, the null hypothesis under both White's (2000) RC test and Hansen's (2005) test states that among competing models, the one with the smallest loss i.e., $L_{m,t}$ is not any better than the losses given by benchmark model i.e., $L_{GM-RV,t}$. However, rejecting the null hypothesis means that at least one model produces smaller forecasting errors compared to the benchmark. The loss functions, $L_{m,t}$ and $L_{GM-RV,t}$ for all the models are taken together via the relative measure as shown in equation (9), i.e., multiple testing is carried out to examine the superiority of the benchmark relative to the best performing model in the competing space or vice a versa.

Using numerical results, Hansen (2005) shows that White's reality check test is conservative at detecting the SPA of competing models relative to the benchmark model forecasts and aggressively favours the null hypothesis (the benchmark forecasts) when the alternate model space contains poor and/or irrelevant forecasts. Hansen articulates this drawback and states that the p-value can be spuriously increased when inferior models are present in the spectrum of models. This is due to the variance of the loss function of a poorly specified m-model $\bar{L}_{m,t}$, which may remain large even after the inclusion of better models. He further proposes two modifications to the original RC test, the first uses studentised test statistic and the second specifies a sample-dependent null distribution for the test statistic.

Hansen (2005) document that the approximate distribution of the Hansen's SPA test-statistic is obtained using the stationary bootstrap of Politis and Romano (1994), similar to RC SPA test-statistic. Nonetheless, the resultant test-statistic has greater power and is less sensitive to poor or irrelevant specifications resulting in the non-rejection of the null hypothesis. Hansen (2005) considered different adjustments to equation (9) to get the three versions of the studentised test statistics that depend on the variance of $\bar{L}_{m,t}$.

4 Data

Our data set contains the market MSCI equity indices (obtained from DataStream) of four major global markets: France, Germany, the UK and the US The daily and monthly dollar prices of these indices are retrieved from Thomson Reuters DataStream for the period from January 1994 to June 2022. We compute continuously compounded returns: $r_i = \ln(P_i) - \ln(P_{i-1})$ for all four equity indices, where *P* represents the daily closing index price levels and *i* is the day index.

Equity markets considered in our study provide consistency and a coherent developed market perspective and therefore, are crucial for global investors, portfolio managers and institutional investors considering the criteria of investibility, replicability and cost efficiency. This is displayed by the fact that MSCI USA alone makes 54% of MSCI all country world index and France, Germany and the UK make up in aggregate 60.72% (17.34%, 15.02% and 28.36%, respectively) of the MSCI Europe index that covers 85 percent of free float capitalization of European equity markets. These indices underline many financial derivative products, exchange traded funds etc. and therefore forecasting its volatility is crucial for pricing, risk management and asset allocation decisions.

[Insert Table 1]

The macroeconomic variables, retrieved at the monthly frequency, for comparable periods are taken from three different sources. The data for France, Germany and the UK are taken from the Eurostat database, whereas the US data come from the Federal Reserve Bank of St. Louis i.e., the FRED database. If the macro data were unavailable from either Eurostat or the FRED, we used the Organization for Economic Co-operation and Development (OECD) database. From these monthly values the growth rate is computed as the natural log difference, except for the term structure of interest rates where we take the simple difference of the yields 10-year maturity bond and monthly T-bills or any monthly duration interest bearing security. To compute the macro-volatility of undertaken aggregate variables in our work, we use innovations from autoregressive-moving average (ARMA) models applied on each macro variable growth series (Schwert 1989).[17]

[17] We use innovations after fitting a best fit ARMA model as given by Bayesian information criterion.

In total, we estimate GM models using eight macroeconomic variables: the term structure of interest rates (TermStr), the exchange rate (EXR), the narrow money (NM), the broad money (BM), the consumer price index (CPI), the industrial production (IP), crude oil prices (Oil) and the unemployment rate (UnEmp).[18] These macro variables are strongly interdependent, especially for developed countries and using them simultaneously, can cause issues pertaining to multi-collinearity, over parameterisation and model convergence.[19] To deal with these aspects, we employ dynamic principal component analysis (PCA) on these eight macro series, for each country, to assimilate information that explains the variance of these variables through meaningful and integrating components.[20] Furthermore, Asgharian et al. (2013) and Virk and Javed (2017) have shown that integrated changes in the macro environment, as captured by the first two components coming from PCA analysis, have clear information benefits relative to GM model that uses a singular macro variable. Therefore, in addition to the eight macro variables, we take the first two components from the PCA analysis for each market when we estimate competing and benchmark models in this study. The summary statistics for the daily returns and squared returns on all four equity indices are presented in Appendix[21] Table A. We also provide the time series patterns of these two during the sample period of work in Appendix C.

[Insert Tables 2 and 3]

5 Empirical results and discussions

We start our estimation period from January 1999 so we can capture sample trends in market volatility.[22] To maximise on changing trends in the period prior to year 1999, we employ MIDAS lags of 5-year[23] duration starting from January 1994. Hence, the one step ahead volatility is forecasted with 5 years of lagged RV and/or macroeconomic data depending on model specification.

[18] The exchange rate for France, Germany and the UK are taken against USD, whereas for the US we take it against a basket of currencies, as provided by FRED.

[19] The issues pertaining to convergence for MIDAS regressions were frequented quite often due to non-convex objective function when we employed flexible weighting in search of optimal weight structure to exploit the information content in each aggregate variable.

[20] PCA analysis has the benefits of over parametrization in model estimations when conditional variance can be influenced by a range of factors and effectively removes noise from the signal. We apply the dynamic PCA such that the stationary macro series are transformed to have standard normal distribution with zero mean and unit variance. We note that the first two components from dynamic PCA invariable explain 70-90 % of the variability in the total factor variance of the macro environment of economies investigated in our work.

[21] All Appendices are provided in the **Supplemental online material** file.

[22] This choice of a 5-year period is to alleviate potential structural breaks that are possible due to the changing patterns in the run to introduction of the Euro, the Global Financial Crisis and the COVID-19 pandemic. These breaks may affect the estimation of conditional volatilities, both total and long-term, when period variations in different macro and market variables may influence their evolution patterns.

[23] Different lag structures are implemented in the estimation process and based on optimal convergence of the model together with minimum use of data for smoothing, 5-year lag (K=5) is selected.

Table 1 shows the set of GM models that are competing against the benchmark GM-RV model. Furthermore, for all models, we estimate two GM models: the first in levels using the growth in the macro variables or the principal components, see equation (6) and the second is when the lagged volatilities of each of the macro variables or the principal components are taken together with the lagged levels, see equation (7). Finally, given the noted gains of principal components in improving the GM forecasting ability in Asgharian et al (2013) and Conrad and Loch (2015), we augment our benchmark model GM-RV with PC1 or PC2 (the first and second principal components obtained through involving all the macro variables), see models 12 and 13 in Table 1 and equation (8) for reference.

In sum, we compare the performance of 26 competing models against our benchmark GM-RV while using two different weighting schemes.[24] That is, all GM models are estimated with unconstrained weights (both the w_1 and w_2 are estimated) or constrained weights (when only w_2 is estimated and w_1 is kept fixed, i.e., $w_1 = 1$) for the beta smoothing function $\phi_k\ (w_1, w_2)$ for each lagged input variable in the MIDAS filter. [25]

5.1 Forecasting errors of the competing models relative to benchmark model

In this section, we compare the relative performance of the quadratic loss functions of all competing models relative to total and secular component forecasts coming from the benchmark GM-RV model. Table 2 presents these ratios using the unconstrained/flexible weighting scheme in the MIDAS filter for all four markets. As noted in Engle et al. (2013) relative performance covers POS forecasts: the parameters for all GM models are estimated using the full sample, and the forecasts for next month are computed using month-end price data.

The MSE ratios under the headings $\sigma^2_{w1,w2}$ and $\tau_{w1,w2}$ are, respectively, for the total variance forecasting errors and the secular components. As described earlier, we estimate two GM models for each MIDAS input variable. Therefore, columns under the heading of X_l describe the ratios of the models that use the level of first order change in the input variables, while the columns under the heading X_{l+v} describe the ratios for models that use both the levels and the volatility of the input variables in the MIDAS filter.

[24] We also compare the MSE forecasting errors of competing errors using another benchmark where we fix the RV in a month i.e. a fixed span GM-RV model. The results of these comparisons are available upon request and qualitatively resemble what we report using the benchmark GM-RV model in this study.

[25] The GM estimations show that, across all models and markets, the estimates in the GARCH model are usually significant at 5% or below critical t-values and estimate values are in line to the vast available evidence for GARCH (1,1) models. The regression coefficients, in all models and across markets, on RV in the MIDAS regressions are positive and super significant regardless of the choice of weighting scheme. However, we note that MIDAS input variables in the competing GM models are more often significant with unrestricted weighting scheme at conventional 5% critical t-values. The significance of these estimates with the restricted weighting scheme reduces drastically. These results are not reported to conserve space for the large number of regressions, with even larger number of regression estimates, carried in our work and available upon request.

The vast majority of the ratios reported in Table 2 across all four markets are less than one indicating enhanced informativeness of the competing forecasts relative to conditional predictions given by the GM-RV model. This result highlights the benefit of incorporating macro information in forecasting total and long run variance components.

However, there are exceptions – a higher quadratic loss is noted in some instances. For France, for both total and long run variance predictions, the MSE ratios are relatively larger when the competing GM model's MIDAS input variables include level and variance of input variables. For the UK and the US, relative rations are near one in more of the instances but preserves the finding that macro variables contain additional information to forecast total and long run variances. Increase in the competing models MSE relative to the benchmark when PC1 and PC2 are included as input variables follows results for France. Results for the US for the models capturing common variation among the macro variables is one exception. The best chance for the benchmark model, as shown by the MSE ratios, is observed for Germany. There are 23 and 16 instances out of 26 MSE ratios the benchmark predictions brought smaller forecasting errors for total variance and long run variance respectively.

Using flexible weighting across the sample countries, we find that term structure of interest rates, oil price changes, unemployment and GM-RV model together with PC1 or PC2 bring valuable forecasting gains relative to the benchmark model predictions. However, gains are far larger for prediction of long run variance component than total variance forecasts. In addition, for secular component forecasts there are other macro variables that suppress forecasting errors better than the benchmark model. Noteworthy results include the relevance of exchange rate changes for the UK and the US total and long run variance forecast. Inflationary pressures are important in suppressing forecast errors for the UK total and long run volatility. However, inflationary changes only bring important information relative to the benchmark for France total volatility and the US secular volatility predictions.

On average, the best models across the three European markets are GM with term structure of interest rates, GM-RV+PC1 and GM-RV+PC2. However, for the US, GM with oil reduces forecasting errors the most relative to the benchmark model.

We replicate the results in Table 2 using the restricted beta smoothing function: $w_1 = 1$, where only w_2 is estimated as a free parameter for each input variable in the MIDAS filter. Table 3 provides the MSE ratios with constrained beta smoothing and the evidence against the benchmark model forecasts is relatively weaker than observed with the flexible weighting scheme in Table 2. One, not all the macro-variables that brought lower forecasting errors using flexible weighting remain relevant as we adopt a restricted weighting scheme, e.g. GM-TermStr for France and the

US. However, in the wake of the cost of living crisis following the COVID-19 pandemic, the restricted weighting in MIDAS smoothing, assigning larger weight to recent observations of the input variable, drives the increased forecasting accuracy for GM-CPI. This finding is displayed by better performance of GM-CPI with restricted weighting in each market than the corresponding results while using flexible weighting scheme. Two, the forecasting accuracy, on average, is lower than what we observed with the flexible weighting scheme. Three, the gains of the macro-information in reducing forecasting errors persists. On average, noticeable total variance and long run variance forecasts are given by GM-RV+PC1 for France, G-RV+PC2 for Germany and GM-TermStr for the UK. For the US, improved forecasting accuracy over the benchmark total variance and secular component are available through GM-UEmp and GM-NM.

[Insert Tables 4 and 5]

We note that including the volatility of the input variables together with the levels bring a more accurate forecast regardless which weighting scheme we apply. This observation persists for all the markets undertaken in our work.

5.2 Pairwise model comparisons: tests of equal forecasting ability

The quadratic losses reported in Tables 2 and 3 are only in-sample estimates and follow Engle et al. (2013) in labelling them as pseudo variance forecasts. Hence, we need to infer the accuracy of the models in the population. We proceed by conducting a pairwise comparison using the DM test, defined in equation (9), to compare the forecasts of the GM-RV benchmark for both total and long-run variances in the set of competing models. These results are reported in Tables 4 and 5.

We note the null of equal predictive ability is rejected across all model specifications and countries. However, the negative test statistic values for the DM test in Table 4 and 5 imply that the competing model outperforms the benchmark model forecasts at conventional 5% or below significance levels. This result reinforces the findings in Tables 2 and 3. Thus, there is overwhelming evidence that the competing models' POS forecasts, for both total and long variances, bring forecasting accuracy. These results mirror image findings with respect to the use of each type of weighting scheme in relation to MSE ratios, as reported in Tables 2 and 3, consistently across all the sample markets.

Given the statistical rigor of the DM test compared to the MSE ratios, the difference in outputs for using flexible and restricted weighting schemes is reinforced: different weighting schemes generate different inferences in terms of forecasting efficiency. We make two interlinked observations. First, the non-EPA of the competing GM models for the total variance forecasts with the restricted weighting scheme is consistent with the results in Engle et al. (2013) and Conrad and Loch (2015): they report that the restricted weighting scheme is optimal for the GM-RV

model. We show superimposing restricted weighting on competing GM models is potentially inapt. Second, this stresses the importance of estimating the best weighting combinations for GM models that incorporate macro variables to let these specifications have a full chance computing their variance forecasts relative to the benchmark forecasts. This is particularly applicable in the evaluation of the total-variance forecasts.

Nonetheless, this does not appear to influence the evidence against the long-run variance forecasts: macro information improves forecasting efficiency of MIDAS component using either weighting scheme. This aspect was previously noted with restricted weighing only in Asgharian et al. (2013).

[Insert Table 6]

5.3 Multiple model comparisons: tests of superior predictive performance

The previous sections have collected evidence that shows how (relatively) weak the benchmark model is: many models outperform the GM-RV model. Invariably, the DM test shows that the inferences from MSE ratios are sensitive to the type of weighting scheme applied, especially for evaluating total variance forecasts. Although our specification search for a good forecasting model is extensive and has samples from four countries, it is possible they are the results of coincidence and data mining. To deal with this, we carry out the reality check (RC) (White, 2000) and superior predictive ability (SPA) tests (Hansen, 2005).

Following the null hypothesis that no model is better than the benchmark model, the RC and SPA tests are not only robust to data snooping issues but are also more powerful in their ability to jointly compare the performance of multiple models. A rejection of null implies that there is at least one model in the competing set of models which is better than the benchmark model.[26]

Before examining the results of multiple testing using RC and Hansen SPA tests, we assess how many of the 26 competing models for each market outperform the benchmark model using the naïve p-values. Note that, the naïve p–values are the bi-model – a variant of DM test – bootstrapped RC test values that are computed as if the best model is the only model in the competing model space. These p-values are an important signal to check for the data snooping bias. White (2000) shows that the naïve p-values are the lower threshold for the RC multiple test p-values: therefore, he suggests it is only meaningful to test for data snooping bias when the p-values are small.

[Insert Tables 7]

As the RC test with naïve p-value below 0.05 implies rejecting null, we count all the instances for variance forecasts (both total and long-run) and divide them by the total number of competing

[26] Both are tests for superior predictive ability as ruled by the null hypotheses of RC/White test and SPA/Hansen test.

models (i.e., 26 in this case). We report these proportions in Table 6 for the POS loss functions. Results show that there at most 2 or 3 models that appear to have to have forecasting efficiency over the benchmark total variance forecast using either of the two weighting schemes. With the restricted weighted scheme the number of models bringing forecasting efficiency is even more curtailed: for the UK and the UK there no models that reduce losses more than the benchmark model forecast.

However, there are relatively more competing GM models that outperform the benchmark model's long-run variance forecasts. We find the largest (lowest) model forecast outperforming the benchmark model for France (the UK) using the restricted weighting scheme i.e. 8 (4). Further, we note that using naïve p-values results, there are more competing GM models with increased forecasting efficiency when unrestricted weighting is used relative to the restricted weighting scheme. This again endorses our expectation that the restricted weighting scheme can undermine the utility of macro information in GM models.

These results, together with the findings reported in sections 5.1 and 5.2, show that there are competing GM models that can be more informative or can reduce MSE more than the benchmark model. However, this inference is weaker when it comes to total variance forecasts consistent with DM test with the restricted weighting scheme. Thus, this observation questions the practice of drawing inferences from simplistic tests combined with a restricted weighting scheme, at least for the total variance comparisons. It also suggests increased caution in evaluating GM forecasts coming from competing models.

Therefore, robust SPA tests that account for multiple testing are critical when carrying out model comparisons to improve inferential reliability. Table 7 shows these results for the RC test and Hansen test. Following Hansen (2005), we compute a test statistic for the consistent (centre, c) bounds for both RC test and Hansen test. Each test-statistic p-value is computed from 1000 bootstrap resamples and bootstrap parameter q= 0.25.[27] The least naïve p-value from all the competing pairwise RC tests is also reported to see the full effect of the joint testing to capture the data mining bias.[28]

The provided p-values for the long-run variance forecast comparison show that the naïve p-values are far lower than the test statistics that employ multiple testing and are thus, biased against the

[27] Our results remain unchanged if we increase the number of bootstrap samples to 10,000. Furthermore, we assess the sensitivity of the SPA results with different values for bootstrap parameter 'q' that controls for time dependence: a q=1 completely ignores the time dependence. The results using lower time dependence than q=0.25 do not alter our results in any manner. We also note that Hansen (2005) and Gonzalez et al. (2004) also use time dependence values of q=0.25 that accounts for substantial time dependence corresponding to the empirical observations in the time series modelling of the equity volatility.

[28] White (2000) reports that the difference between the naïve p-value and RC test can be described as the data snooping bias in the specification search of better models than the benchmark model.

null. The p-values for both the test statistics i.e. RC_C, and $Hansen_C$, are fairly large and do not reject the null – supporting the total variance forecast accuracy over the competing model forecast space. In other words, the benchmark model's total variance forecast cannot be outperformed given the *RV* based approximation of unobservable daily volatility. This is a startling inference which, when linked with the results given by the DM test, especially with flexible weighting, shows that the results from even the DM test are untenable when the data snooping bias is considered.

However, the picture is slightly different in the case of long run volatility forecasts. The joint testing substantiates the conjecture that at least one of the competing models, which incorporate economic information, generates a more accurate long-term volatility forecasts than benchmark. The null hypothesis is rejected by both RC and Hansen SPA tests when the flexible weighting scheme is employed.

For the restricted weighting scheme, the same inference follows and the only exception is the US results for variance secular component forecast. For the US, the benchmark model's prediction for the long-run volatility shows superior predictive ability: the naïve p-value for the best performing model is 0.078 and as we have observed in the rest of the instances RC and Hansen p-values are far greater than the naïve p-values.The latter observation is in line with Hansen (2005) the SPA p-values are relatively lower than the corresponding RC test statistic values. For example, in the US secular variance forecast comparisons when restricted weighting is used RC_c is 0.531 while $Hansen_c$ p-value is 0.478.[29] For the US data, our results are in accordance with findings reported in Conrad and Kleen (2020) that the GM-RV model is outperformed by models involving macroeconomic information.

Overall, we infer that GM models incorporating macroeconomic information may bring benefits in forecasting the long-run variances but the same cannot be concluded with respect to their total variance forecasts.

[Insert Table 8]

5.4 Out-of-sample model forecast comparisons

Until now we have carried out all tests on the pseudo out-of-sample forecasts (Engle et al. 2013, Conrad and Loch 2015). However, to test the full effect of our results reported in Table 7, we compare the out-of-sample (OS) forecasts of all the competing models using a rolling window

[29] Hansen (2005) shows that RC tests can be manipulated in the presence of poor and irrelevant alternatives in the sample of models. This results in less power and non-rejection of the null hypothesis. Hansen (2005) alleviated this problem in his version of the SPA test by invoking a sample-dependent distribution under the null hypothesis.

forecasting scheme.[30] To conduct the out of sample forecasts comparison, we get OS conditional total variance and long-run variance forecasts for 2012-16 for one year at a time.

[Insert Tables 9]

Procedurally, the parameters are obtained using a rolling 13-year estimation window for each GM model, which are held constant during the subsequent year to compute daily (252 step-ahead) total variance and monthly (12 step-ahead) long-run variance forecasts. The estimation sample is then moved one year ahead to re-estimate the parameters and the forecasts are made for the next subsequent year and so on. Thus, we perform this procedure five times to obtain total and secular variance forecasts for each year from 2012 to 2022. The estimation windows correspondingly cover rolling sample periods of 1999-2011, 2000-2012, …, 2010-2022.

We present the p-values for the RC and Hansen tests in Table 8 for the out of sample forecasts. Table 8 shows that using the flexible weighting scheme the RC tests have large p-values and null cannot be rejected at 0.05 significance p-values for the total variance forecast errors for any market. On the contrary when it comes to the $Hansen_c$ statistic: the SPA of the benchmark total variance forecast is rejected for all except for the US. With the restricted weighting scheme, both tests reject the notion that any competing models' forecasts have efficiency gains in forecasting total variance relative to the benchmark model forecast with cut-off of 0.05 p-values for the data-snooping test statistics.

The OS forecasting comparisons for the long-run variance forecasts show that using either weighting scheme RC_c and $Hansen_c$ statistic p-values reject the null for all the markets. Further, we note that the data-snooping bias is sizeable as the difference between naïve p-values and RC_c or $Hansen_c$ are large.

In order to check the sensitivity of the results reported in Tables 7 and 8, we replace RV with two other variance proxies, i.e. conditional variance forecasts from GARCH (1, 1) model and the variance of implied volatility indices of the markets undertaken in this study. We present the OS outputs in Table 9.[31] The results show that with either of the proxies used, forecasting gains with aggregate variables are present for the UK and the US. The same is not the case for French and German market volatility resorting to $Hansen_c$ statistic p-values while using flexible weighting scheme. When it comes to long run variance forecasts, benchmark forecast cannot be outperformed for German, the UK and the US market total variances with GARCH (1,1) proxy.

[30] Multiple steps ahead forecasts for long periods such as five years are not feasible in the context of the models examined in our work, knowing latent variance is time varying and the expected future value of the macroeconomic growth and volatility is not measurable given the time of forecast information filter.

[31] Results for POS forecast are available upon request and they largely follow the same pattern that we observe for OS forecasting comparisons.

Using implied volatilities of each market only returns forecasting gains for German and the US market.

The analysis of the restricted weighting scheme is presented in the right panel of Table 9. The forecasting gains are only observed for the German long run variance with implied volatility as the proxy for latent variance component. The forecasting gains for France, the UK and the US are absent using models that incorporate macro-variables.

5.5 Additional Tests

Following Engle et al. (2013), we run variance ratio tests. These ratios mainly depict the differences in the output as models that provide larger weight to recent observations – the restricted weighted scheme – mostly result in poorer forecasts for the secular component of the total variance in contrast to forecasts based on the flexible weighting scheme. These findings remain consistent whether we compute ratios using the model's own total variance or the benchmark model's total variance. For details refer to Tables B1 to B4 in Appendix B of the supplementary results.

We examine the predictive ability of long-term variance from GM models for log realised volatility. Similar regressions are also reported in Asgharian et al. (2013) and Conrad and Loch (2015). These predictive regressions endorse our earlier evidence supporting the choice of the unrestricted weighting scheme relative to the restricted weighting scheme. For details on these regressions please refer to Tables B5 and B6 in Appendix B provided in the supplementary files.

6 Summary of Results and Conclusions

To summarise, we divide our results in two parts for ease of description. In the first part, we replicate evidence from Engle et al. (2013), Asgharian et al. (2013) and Conrad and Loch (2015) using pairwise MSE ratios, variance ratios, DM test and predictive regressions. In this respect, we provide new results using French, German and British financial and economic datasets. In the second part, we complement new evidence in the GM variance forecasting literature by examining data snooping issues in making the forecast comparisons among GM models. Our results from the first part show that the MSE of the GM models that include macro information, alone or with the RV in the MIDAS regressions and with either type of weighting scheme, are smaller than the benchmark model. These findings are consistent with earlier studies. We find that total and long-run variance forecasts from the GM model that use RV and PC1 give the best performance across all markets when we use a flexible weighing scheme. While results using the restricted weighting scheme find even better results when using macro information in the GM models. These findings

are also confirmed by the EPA DM test – only problem found is the use of restricted weight in forecasting total variance. That is, the DM tests provide overwhelming evidence that the competing models POS forecasts, for both variance components, bring efficiency gains while using the flexible weighting scheme. These results are uniform across the four countries. However, when we adopt the restricted weighted scheme the consistency in results is only witnessed for the long-run variance forecasts.

Drawing together the results of MSE ratios and the DM test, we note that there are differences in forecasting gains while using a particular weighting scheme in the MIDAS smoothing using either type of tests. However, the use of informative ratios in making reliable inferences is untenable. Furthermore, the DM test illustrates that the benchmark total variance forecasts are only as good as few of the competing model forecasts using a flexible/unrestricted. These instances reduce when a restricted weighting scheme is employed. This difference is starker when we compare the forecasting gains for the long run variance using the weighting schemes. This observation implies that projecting a weighting structure that is optimal for the GM-RV model may result in limiting the scope for the macro variables in the GM models to predict the secular variance component. Thus, we suggest that econometricians and practitioners should be wary of this simplification while searching for precise and reliable variance forecasts. Essentially, superimposing a particular evolutionary structure in MIDAS may yield results that favour the null even when that is not the case. Therefore, incorporating relevant low frequency information, and the consequent evolution and approximation of variances, should not be comprised by resorting to a general weighting pattern that might not be suitable for the sake of simplicity.

The low reliability of these results when assessed through powerful tests that account for data mining endorse our scepticism: joint testing shows that no model outperforms the benchmark model's total variance forecast in all markets. However, evaluation of the long-run variance coming from competing GM macro models there are clear gains over the benchmark model's secular trend. With respect to multiple testing, we note the value of pairwise tests: if the p-value for DM test or naïve robust test is large, there is no need for data snooping bias testing because, as White (2000) reports, p-values from multiple tests can only be larger than the naïve p-value. Overall, the generality of our main results with respect to scepticism to data snooping biases and adoption of a particular weighting scheme persists for using different proxies for latent variance and using a rolling OS forecasting scheme.

We summarise our results by making the following four observations. First, our results show that inferences based on informative or pairwise model comparison tests can be misleading, especially when they are about the gain from using macro information in projecting total GM variance. This

is especially dangerous when we have seen the evolving evidence studying GM models that have generalised the evidence for total variance to long-run variance forecasts. Second, we show that GM models are an important addition in forecasting volatility at different time frequencies, but we have to be cognisant of what type of variance forecast we are looking for. Third, we should not compromise the weighting scheme for the MIDAS input variables by making ad hoc choices and, fourth, we should be wary for data mining biases that may plague a forecast that otherwise appear statistically validated.

These results are important given the demand for reliable, accurate variance forecasts for devising risk planning and management protocols at institutional and individual levels, especially since investors vary in their investment choices across time horizons, e.g., active and passive investment decisions require variance forecasts for two different investing styles. Therefore, using macro variables adds information that improve modelling and forecasting of long-run or total variance, we should capitalise on it. This applies when there are several low frequency processes which contain the relevant, independent information needed to forecast latent variance for different time horizons. For future research, we recommend use of high frequency data in the GM modelling to develop precise proxies for latent variance while linking it to different dimensions of sentiment as well as economic and political uncertainty indices.

Acknowledgements

The corresponding author acknowledge financial support from the project "Models for macro and financial economics after the financial crisis" (Dnr: P18-0201) funded by the Jan Wallander and Tom Hedelius Foundation.

References

Andersen TG, Bollerslev T, Diebold FX, Ebens H., The distribution of realized stock return volatility. *Journal of Financial Economics,* 2001b, **61**, 43-76.

Andersen, T. G. and Bollerslev, T., Answering the Skeptics: Yes, Standard Volatility Models do Provide Accurate Forecasts, *International Economic Review,* 1998, **39** (4), 885-905.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P., The distribution of realized exchange rate volatility. *Journal of the American statistical association*, 2001a, ***96***(453), 42-55.

Andersen, T.G. and Bollerslev, T., Intraday Periodicity and Volatility Persistence in Financial Markets, *Journal of Empirical Finance,* 1997, **4**, 115-158.

Asgharian, H., Christiansen, C. Hou, A., Macro-Finance Determinants of the Long-Run Stock–Bond Correlation: The DCC-MIDAS Specification, *Journal of Financial Econometrics*, 2016, **14** (3), 617–642.

Asgharian, H., Hou, A., Javed, F., The importance of the economic variables in predicting the long term volatility; a GARCH MIDAS approach, *Journal of Forecasting*, 2013, **32**, 600-612.

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

Barndorff-Nielsen, O. E., & Shephard, N., Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society*: Series B (Statistical Methodology), 2001, **63**(2), 167-241.

Christie, Andrew A., The stochastic behavior of common stock variances: Value, leverage and interest rate effects, *Journal of Financial Economics*, 1982, **10**, 407-432.

Conrad, C., Kleen, O., Two are better than one: Volatility forecasting using multiplicative component GARCH-MIDAS models. *Journal of Applied Econometrics*, 2020, **35**, 19-45.

Conrad, C., Loch, K., and Rittler, D., On the macroeconomic determinants of long-term volatilities and correlations in U.S. stock and crude oil markets, *Journal of Empirical Finance*, 2014, **29**, 26-40.

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

Cont, R., Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative. Finance*, 2015, **1**, 223-236.

Degiannakis, S., Filis, G. and Kizys, R., 2014. The effects of oil price shocks on stock market volatility: Evidence from European data. The Energy Journal, 35(1).

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

Eichengreen, B. and Tong, H., 2003. Stock market volatility and monetary policy: what the historical record shows. Asset Prices and Monetary Policy, pp.108-42.

Engle, R. F., Ghysels, E., & Sohn, B., Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics*, 2013, ***95***(3), 776-797.

Fang, L., Yu, H., and Huang, Y., The role of investor sentiment in the long-term correlation between U.S. stock and bond markets, *International Review of Economics & Finance*, 2018, **58**, 27-139,

Ghysels E., Santa-Clara P., Valkanov R., Predicting Volatility: Getting the Most out of Return Data Sampled at Different Frequencies, *Journal of Econometrics*, 2006, **131**, 59-95.

Ghysels, E., Santa-Clara, P., and Valkanov, R., The MIDAS touch: Mixed Data Sampling Regression. Discussion Paper UNC and UCLA, 2004.

Ghysels, E., Sinko, A., and Valkanov R., MIDAS Regressions: Further Results and New Directions, Econometric Reviews, 2007, **26**, 53-90.

González-Rivera, G., Lee, T. H., & Mishra, S., Forecasting volatility: A reality check based on option pricing, utility function, value-at-risk, and predictive likelihood. *International Journal of forecasting*, 2004, ***20***(4), 629-645.

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

Hansen, P. R., A test for superior predictive ability. *Journal of Business & Economic Statistics*, 2005, ***23***(4), 365-380.

Lindblad, A., Sentiment indicators and macroeconomic data as drivers for low-frequency stock market volatility. MPRA Paper 80266, University Library of Munich, Germany, 2017.

Meddahi, N., A theoretical comparison between integrated and realized volatility. *Journal of Applied Econometrics*, 2002, **17**(5), 479-508.

Officer, R. R., The variability of the market factor of the New York Stock Exchange. *The Journal of Business*, 1973, **46** (3), 434-453.

Pan, Z., Bu, R., and Wang, Y., Macroeconomic fundamentals, jump dynamics and expected volatility, *Quantitative Finance*, 2020, **8**, 1345-1371.

Pan, Z., Wang, Y., Wu, C., and Libo, Y., Oil price volatility and macroeconomic fundamentals: A regime switching GARCH-MIDAS model, *Journal of Empirical Finance*, 2017, **43**, 130-142.

Patton, A. J., Volatility forecast comparison using imperfect volatility proxies. Journal of Econometrics, 2011, **160**, 246-256.

Politis, D. N., & Romano, J. P., The stationary bootstrap. *Journal of the American Statistical Association*, 1994, **89**, 1303-1313.

Schwert, G. W., Why does stock market volatility change over time? *Journal of Finance*, 1989, **44**(5), 1115-1153.

Shiller, Robert J. Alternative tests of rational expectations models: The case of the term structure. *Journal of Econometrics,* 1981a, **16**, 71-87.

Shiller, Robert J., The use of volatility measures in assessing market efficiency, *Journal of Finance*, 1981b, **36**, 291-304.

Virk, N. and Javed, F., European equity market integration and joint relationship of conditional volatility and correlations, *Journal of International Money and Finance*, 2017, **71**, 53-77.
West, K., Asymptotic inference about predictive ability. *Econometrica*, 1996, **64**, 1067-1084.
White, H. A reality check for data snooping. *Econometrica*, 2000, **68**(5), 1097-1126.

Table 1: The Set of Competing GARCH-MIDAS Macro Models

The first line of the table shows the table benchmark model used in this study i.e., GARCH-MIDAS (GM) model that includes the monthly rolling window (RW) realized variance (RV) in the MIDAS filter. Models 1-11 only include the level of the listed macro variable in the GM model. The last two models, i.e., 12 and 13, add the level of the first and second principal components of all the macro variables shown in the table to the benchmark GM-RV model. We estimate 13 additional models that include both the level and the volatility of the explanatory variables in each of 13 competing models shown below. Hence, a set of 26 models compete against the benchmark GM-RV.

Model	**GARCH-MIDAS models**	**Acronym**
Benchmark model	realized variance (RV)	GM-RV
1	The Term Structure of Interest rates	GM-TermStr
2	Exchange Rate	GM-EXR
3	Narrow Money	GM-NM
4	Broad Money	GM-BM
5	Consumer Price Index	GM-CPI
6	Industrial Production	GM-IP
7	Crude Oil Prices	GM-Oil
8	Unemployment	GM-UEmp
9	First Principal Component, PC1	GM-PC1
10	Second Principal Component, PC2	GM-PC2
11	With PC1 and PC2	GM-PC1+PC2
12	RV with PC1	GM-RV+PC1
13	RV with PC2	GM-RV+PC2

Table 2 MSE Ratios using flexible weighting scheme in the GM models

In all the estimated models, the MIDAS explanatory variables span the past five-year lagged data (from January 1994 to December 1998). The models use monthly input variables except the benchmark model where we use monthly realized variance – 22 day rolling window – that is available daily. Thus, the two-component volatility models are fitted over the period of January 1999 to June 2022 to carry out pseudo out-of-sample forecast comparisons using MSE ratios of competing model relative to GM-RV model i.e., the benchmark model. Results are presented in blocks for France, Germany, the UK and the USA. Here $\sigma_{\boldsymbol{w}}^2$ and $\tau_{\boldsymbol{w}}$ refer to the comparisons of total daily volatility and monthly secular volatility from competing k-models with the GM-RV model. The full sample parameters for each model are subsequently taken to compute the next period pseudo forecasts. The MSE of the forecasts given by each model is thus calculated, including the benchmark model, with respect to monthly RV. For each equity market, MSE ratios below $\sigma_{w1,w2}^2$ and $\tau_{w1,w2}$ are computed when the parameters of the beta polynomial function are unconstrained (i.e., w_1 and w_2 for each input variable are allowed to be free parameters). The competing model is determined by the interaction of the models below the heading "competing models" with column headings X_l and X_{l+v}. This implies that the MSE ratios are separated across *k*-models that use only the level of the MIDAS input variable i.e., X_l and the ones that use the level and volatility of the input variables combined in the MIDAS filter i.e., X_{l+v}. All pseudo-forecasts are generated using fixed full sample parameter estimates as in Engle et al. (2013).

	France				Germany			
	$\sigma_{w1,w2}^2$		$\tau_{w1,w2}$		$\sigma_{w1,w2}^2$		$\tau_{w1,w2}$	
Competing models	X_l	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}
GM-TermStr	0.564	0.637	0.282	0.092	1.026	0.958	0.522	0.522
GM-EXR	0.496	0.458	0.304	0.142	1.109	1.123	1.284	1.287
GM-NM	0.334	1.016	0.251	1.001	1.039	1.098	0.317	1.288
GM-BM	0.900	0.544	0.999	0.647	1.089	1.105	1.284	1.284
GM-CPI	0.935	0.917	1.000	0.999	1.093	1.035	1.283	1.283
GM-IP	0.783	1.100	0.998	0.999	1.021	0.994	1.282	1.282
GM-Oil	0.512	0.677	0.465	0.974	0.965	1.063	0.218	1.285
GM-UEmp	0.593	0.540	0.440	0.085	1.040	1.027	0.621	0.621
GM-PC1	0.822	0.867	0.998	0.998	1.056	1.003	1.283	0.310
GM-PC2	0.939	0.925	0.999	1.001	1.113	1.037	1.285	1.285
GM-PC1+PC2	0.916	1.083	1.000	1.001	1.059	1.064	1.282	1.283
GM-RV+PC1	0.929	0.206	1.000	0.015	1.075	1.038	1.285	0.201
GM-RV+PC2	1.072	1.105	1.000	1.002	1.048	1.064	0.437	0.554
	UK				USA			
	$\sigma_{w1,w2}^2$		$\tau_{w1,w2}$		$\sigma_{w1,w2}^2$		$\tau_{w1,w2}$	
Competing models	X_l	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}
GM-TermStr	0.797	1.000	0.310	1.000	0.982	0.963	0.374	0.656
GM-EXR	0.995	0.816	1.000	0.612	0.998	0.987	0.999	0.999
GM-NM	0.897	1.012	0.513	0.999	0.995	1.002	1.000	0.999
GM-BM	0.871	0.889	0.169	0.302	0.999	0.965	1.000	0.375
GM-CPI	0.897	0.899	0.318	0.999	1.003	1.000	0.998	0.998
GM-IP	0.994	0.993	0.999	1.000	0.996	1.003	0.998	0.997
GM-Oil	1.001	0.894	1.003	0.479	0.939	0.954	0.443	0.047
GM-UEmp	0.800	1.008	0.157	0.181	0.971	0.940	0.333	0.032
GM-PC1	0.999	0.998	1.000	1.002	1.005	1.008	0.998	0.998
GM-PC2	0.887	0.999	0.263	1.001	1.003	0.983	0.999	0.999
GM-PC1+PC2	0.792	1.001	0.425	1.001	1.013	0.963	0.999	0.061
GM-RV+PC1	1.001	0.995	1.001	1.003	0.986	1.005	0.208	0.264

GM-RV+PC2	1.037	0.802	0.333	0.296	0.939	0.983	0.264	1.000

Table 3 MSE Ratios using constrained weighting scheme in the GM models

In all the estimated models, the MIDAS explanatory variables span the past five-year lagged data (from January 1994 to December 1998). The models use monthly input variables except the benchmark model where we use monthly realized variance – 22 day rolling window – that is available daily. Thus, the two-component volatility models are fitted over the period of January 1999 to December 2016 to carry out pseudo out-of-sample forecast comparisons using MSE ratios of competing model relative to GM-RV model i.e., the benchmark model. Results are presented in blocks for France, Germany, the UK and the USA. Here σ_w^2 and τ_w refer to the comparisons of total daily volatility and monthly secular volatility from competing k-models with the GM-RV model. The full sample parameters for each model are subsequently taken to compute the next period pseudo forecasts. The MSE of the forecasts given by each model is thus calculated, including the benchmark model, with respect to monthly RV – rolling window or monthly sum- as applicable to the sample of all models. For each equity market, MSE ratios below σ_{w2}^2 and τ_{w2} are computed when the parameters of the beta polynomial function are constrained: weighting scheme fixes w_1= 1 and w_2 for each input variable in the MIDAS filter is estimated as a free parameter. The competing model is determined by the interaction of the models below the heading "competing models" with column headings X_l and X_{l+v}. This implies that the MSE ratios are separated across *k*-models that use only the level of the MIDAS input variable i.e., X_l and the ones that use the level and volatility of the input variables combined in the MIDAS filter i.e., X_{l+v}. All pseudo-forecasts are generated using fixed full sample parameter estimates as in Engle et al. (2013).

	France				Germany			
Competing models	σ_{w2}^2		τ_{w2}		σ_{w2}^2		τ_{w2}	
	$\boldsymbol{X_l}$	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}
GM-TermStr	0.979	1.118	1.000	1.000	0.998	0.891	0.999	0.032
GM-EXR	1.055	0.983	1.000	1.001	1.029	0.941	0.998	0.045
GM-NM	1.115	0.421	1.000	0.054	0.933	1.000	0.246	0.998
GM-BM	0.972	0.816	1.000	1.000	0.994	1.016	0.998	0.999
GM-CPI	0.604	0.766	0.521	0.128	0.992	0.992	0.998	0.998
GM-IP	0.399	1.279	0.023	1.001	1.011	1.024	0.999	0.999
GM-Oil	0.977	0.783	1.000	0.132	1.008	0.997	0.998	0.999
GM-UEmp	0.635	0.783	0.536	0.007	0.881	0.983	0.197	0.998
GM-PC1	0.750	0.970	0.114	1.002	1.008	0.998	0.998	0.999
GM-PC2	1.000	1.217	1.000	1.002	1.002	1.016	0.999	0.999
GM-PC1+PC2	0.768	0.651	0.229	0.230	0.997	1.023	0.999	0.999
GM-RV+PC1	0.742	0.218	0.437	0.016	1.031	1.000	0.998	0.999
GM-RV+PC2	0.681	1.190	0.413	1.002	0.944	0.944	0.278	0.058
	UK				**USA**			
Competing models	σ_{w2}^2		τ_{w2}		σ_{w2}^2		τ_{w2}	
	X_l	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}
GM-TermStr	0.998	1.013	0.404	0.022	1.008	0.972	1.000	0.128
GM-EXR	1.068	1.251	0.281	1.000	0.987	0.961	0.998	0.068
GM-NM	1.228	1.259	0.435	0.999	0.977	0.963	0.004	0.005
GM-BM	1.221	1.032	0.695	0.094	0.994	0.963	0.998	0.897
GM-CPI	0.991	1.135	0.154	0.060	1.000	0.988	0.999	0.999
GM-IP	1.105	1.243	0.085	1.000	0.962	0.993	0.074	0.941
GM-Oil	0.991	1.246	0.127	1.000	0.986	0.966	0.998	1.030
GM-UEmp	1.254	1.252	0.999	1.000	0.913	0.968	0.071	1.066
GM-PC1	1.255	1.311	1.001	0.988	0.990	0.982	0.998	0.074
GM-PC2	1.182	1.082	0.373	0.157	0.974	1.001	0.163	0.999
GM-PC1+PC2	1.007	1.256	0.419	0.999	1.008	0.961	0.999	0.046

GM-RV+PC1	1.264	0.982	0.510	0.186	0.981	1.001	0.147	1.000
GM-RV+PC2	1.292	1.250	0.407	0.999	1.012	0.962	1.000	0.026

Table 4 The Diebold Mariano test using unconstrained weighting scheme

The table presents the t statistics of the Diebold Mariano (1995) test of equal predictive accuracy between pseudo out of sample forecasts from the GM-RV model and the competing k-models. The competing model is determined by the interaction of the models below the heading "competing models" with column headings X_l and X_{l+v}. Here the GM model below column X_l refers to the specification using the levels of the MIDAS input variables only, while X_{l+v} refers to the GM model that uses the level of the input variable together with its volatility in the MIDAS filter. The $\sigma^2_{w1,w2}$ refers to testing the predictive ability of the total daily volatility forecasts of the macro model while the $\tau_{w1,w2}$refers to testing the accuracy of monthly volatility forecasts. This implies that MIDAS beta polynomial across all models is estimated unconstrained. The test-statistic value significant at 0.05 or below p-values are presented with asterisks.

	France				Germany			
	$\sigma^2_{w1,w2}$		$\tau_{w1,w2}$		$\sigma^2_{w1,w2}$		$\tau_{w1,w2}$	
Competing models	X_l	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}
GM-TermStr	-*	-*	-*	-*	+*	-*	-*	-*
GM-EXR	-*	-*	-*	-*	+*	+*	+*	+*
GM-NM	-*	+*	-*	+*	+*	+*	-*	+*
GM-BM	-*	-*	-*	-*	+*	+*	+*	+*
GM-CPI	-*	-*	+*	-*	+*	+*	+*	+*
GM-IP	-*	+*	-*	-*	+*	+*	+*	+*
GM-Oil	-*	-*	-*	-*	-*	+*	-*	-*
GM-UEmp	-*	-*	-*	-*	+*	+*	-*	-*
GM-PC1	-*	-*	-*	-*	+*	+*	+*	-*
GM-PC2	-*	-*	-*	+*	+*	+*	+*	+*
GM-PC1+PC2	-*	+*	-*	+*	+*	+*	+*	+*
GM-RV+PC1	-*	-*	+*	-*	+*	+*	+*	-*
GM-RV+PC2	+*	+*	+*	+*	+*	+*	-*	-*
	UK				USA			
	$\sigma^2_{w1,w2}$		$\tau_{w1,w2}$		$\sigma^2_{w1,w2}$		$\tau_{w1,w2}$	
Competing models	X_l	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}
GM-TermStr	-*	+*	-*	+*	-*	-*	-*	-*
GM-EXR	-*	-*	+*	-*	-*	-*	-*	-*
GM-NM	-*	+*	-*	-*	-*	+*	+*	+*
GM-BM	-*	-*	-*	-*	+*	+*	-*	-*
GM-CPI	-*	-*	-*	-*	+*	+*	-*	-*
GM-IP	-*	-*	-*	+*	-*	+*	-*	-*
GM-Oil	+*	-*	+*	-*	-*	-*	-*	-*
GM-UEmp	-*	+*	-*	-*	-*	-*	-*	-*
GM-PC1	-*	-*	+*	+*	+*	+*	-*	-*
GM-PC2	-*	-*	-*	+*	+*	-*	-*	-*
GM-PC1+PC2	-*	+*	-*	+*	+*	-*	-*	-*
GM-RV+PC1	+*	-*	+*	+*	-*	+*	-*	-*
GM-RV+PC2	+*	-*	-*	-*	-*	-*	-*	+*

Table 5 The Diebold Mariano test using constrained weighting scheme

The table presents the t statistics of the Diebold Mariano (1995) test of equal predictive accuracy between the GM-RV model and the competing k-models. The competing model is determined by the interaction of the models below the heading "competing models" with column headings X_l and X_{l+v}. Here the GM model below column X_l refers to the specification using the levels of the MIDAS input variables only, while X_{l+v} refers to the GM model that uses the level of the input variable together with its volatility in the MIDAs filter. The σ_{w2}^2 refers to testing the predictive ability of the total daily volatility forecasts of the macro model while the τ_{w2}refers to testing the accuracy of monthly volatility forecasts. This implies that MIDAS beta polynomial across all models is where weighting scheme fixes w_1= 1 and w_2 is estimated as a free parameter. The test-statistic value significant at 0.05 or below p-values are presented with asterisks.

	France				Germany			
Competing models	σ_{w2}^2		τ_{w2}		σ_{w2}^2		τ_{w2}	
	X_l	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}
GM-TermStr	-*	+*	+*	+*	-*	-*	-*	-*
GM-EXR	+*	-*	+*	+*	+*	-*	-*	-*
GM-NM	+*	-*	+*	-*	-*	+*	-*	-*
GM-BM	-*	-*	+*	+*	-*	+*	-*	-*
GM-CPI	-*	-*	-*	-*	-*	-*	-*	-*
GM-IP	-*	+*	-*	+*	+*	+*	-*	-*
GM-Oil	-*	-*	+*	-*	+*	-*	-*	-*
GM-UEmp	-*	-*	-*	-*	-*	-*	-*	-*
GM-PC1	-*	-*	-*	+*	+*	-*	-*	-*
GM-PC2	+*	+*	+*	+*	+*	+*	-*	-*
GM-PC1+PC2	-*	-*	-*	-*	-*	+*	-*	-*
GM-RV+PC1	-*	-*	-*	-*	+*	+*	-*	-*
GM-RV+PC2	-*	-*	-*	+*	-*	-*	-*	-*
	UK				USA			
Competing models	σ_{w2}^2		τ_{w2}		σ_{w2}^2		τ_{w2}	
	X_l	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}	X_l	X_{l+v}
GM-TermStr	-*	+*	-*	-*	+*	-*	+*	-*
GM-EXR	+*	+*	-*	+*	-*	-*	-*	-*
GM-NM	+*	+*	-*	-*	-*	-*	-*	-*
GM-BM	+*	+*	-*	-*	-*	-*	-*	-*
GM-CPI	-*	+*	-*	-*	+*	-*	-*	-*
GM-IP	+*	+*	-*	+*	-*	-*	-*	-*
GM-Oil	+*	+*	-*	+*	-*	-*	-*	-*
GM-UEmp	+*	+*	-*	+*	-*	-*	-*	+*
GM-PC1	+*	+*	+*	-*	-*	-*	-*	+*
GM-PC2	+*	+*	-*	-*	-*	+*	-*	-*
GM-PC1+PC2	+*	+*	-*	-*	+*	-*	-*	-*
GM-RV+PC1	+*	-*	-*	-*	-*	+*	-*	+*
GM-RV+PC2	+*	+*	-*	-*	+*	-*	+*	-*

Table 6 Robust pairwise testing for pseudo out-of-sample MSE losses

In this table, we report the proportion of variance forecasts – total and long run – that outperform the benchmark model's corresponding forecasts using the naïve p-values. The naïve p–values are the bi-model – a variant of DM test – bootstrapped RC test values that are computed as if the best model is the only model in the competing model space. Thus, we count all the instances for variance forecasts and divide them by the total number of competing models (i.e., 26 in this case). To exemplify, a proportion of 0 implies that there is no model that outperforms the forecast given by the benchmark model and a proportion of 1 will imply all models from competing model space outperform the benchmark model's prediction. The proportions below the header $\sigma^2_{w1,w2}$ and $\tau_{w1,w2}$ are for the total variance and the secular variance while employing an unconstrained weighting scheme, whereas below headers σ^2_{w2} and τ_{w2} the same is provided while using a constrained weighting scheme in the MIDAS smoothing.

	$\sigma^2_{w1,w2}$	$\tau_{w1,w2}$	σ^2_{w2}	τ_{w2}
	RC	RC	RC	RC
France	0.038	0.346	0.077	0.308
Germany	0.115	0.154	0.038	0.192
The UK	0.077	0.308	0.000	0.154
The US	0.038	0.269	0.000	0.231

Table 7 Multiple testing for pseudo out-of-sample MSE losses

The table presents the p-values of the White's (2000) reality check (RC) test and the Hansen's (2005) superior predictive ability (SPA) test. The forecasting errors i.e., MSE of the competing models are compared against the GM-RV benchmark model using the pseudo put-of-sample model forecasts following Engle et al. (2013). The sample period is January 1999- December 2016. Panels A, B present results for flexible and restricted weighting schemes respectively. First, we present naïve p-values that are the bootstrap RC p-values for the bi-model predictive ability test: bootstrapped p-values by comparing forecasting errors of the best model among $k = 26$ competing models relative to the benchmark model. Following Hansen, we compute the test statistics for the consistent (centre, c), bound for both RC test and Hansen test. The test-statistic p-value is computed from 1000 bootstraps resamples and smoothing parameter q = 0.25. For clarity a p-value larger than 0.05 shows that no competing model outperforms the benchmark model i.e., null hypothesis cannot be rejected, whereas value <0.05 shows that at least one model has superior predictive ability (lower MSE or forecasting errors) than the benchmark GM-RV model, i.e., null is rejected. These tests are carried out using daily squared returns and monthly RV as the latent variance proxies to calculate the loss functions for the full set of models across the four markets for the 26 competing models and the benchmark model GM-RV for each equity market index.

	$\sigma^2_{w1,w2}$				$\tau_{w1,w2}$			
RP metric	France	Germany	UK	US	France	Germany	UK	US
Panel A: flexible weighting scheme								
Naive	0.173	0.096	0.153	0.107	0.000	0.000	0.000	0.005
RC_c	0.457	0.410	0.394	0.472	0.041	0.029	0.031	0.007
$Hansen_c$	0.291	0.357	0.413	0.349	0.008	0.021	0.016	0.01
Panel B: restricted weighting scheme								
	σ^2_{w2}				τ_{w2}			
Naive	0.094	0.157	0.194	0.06	0.000	0.000	0.013	0.078
RC_c	0.239	0.654	0.489	0.519	0.008	0.021	0.041	0.531
$Hansen_c$	0.190	0.429	0.283	0.427	<0.001	<0.001	0.037	0.478

Table 8 Multiple testing for out-of-sample MSE losses

The table presents the p-values of the White's (2000) reality check (RC) test and the Hansen's (2005) superior predictive ability (SPA) test. The forecasting errors i.e., MSE of the competing models are compared against the GM-RV benchmark model using a fixed out-of-sample forecasting scheme. Parameters are estimated using for sample period January 1999- December 2011 and forecasting period is January 2012-December 2016. Panels A, B present results for flexible and restricted weighting schemes respectively. First, we present naïve p-values that are the bootstrap RC p-values for bi-model predictive ability test: bootstrapped p-values by comparing forecasting errors of the best model among $k = 26$ competing models relative to the benchmark model. Following Hansen (2005), we compute three test statistics for the consistent (centre, c) bound for both RC test and Hansen test. The test-statistic p-value is computed from 1000 bootstraps resamples and smoothing parameter q = 0.25. For clarity a p-value larger than 0.05 shows that no competing model outperforms the benchmark model i.e., null hypothesis cannot be rejected, whereas value <0.05 shows that at least there is one model that has superior predictive ability (lower MSE or forecasting errors) than the benchmark GM-RV model i.e., null is rejected. The $\sigma^2_{w1,w2}$ and the $\tau_{w1,w2}$ refers to the results of the forecasting errors of daily total and monthly long run volatilities when the weights for the beta polynomial for each input variable in the MIDAS filter are estimated, while σ^2_{w2} and $\tau_{,w2}$ refers to the model comparison results when $w_1 = 1$ and w_2 is estimated for the beta polynomial of each input variable in the MIDAS filter, respectively. These tests are carried out using daily squared returns and monthly RV as the latent variance proxies to calculate the loss functions for the full set of models across the four markets for the 26 competing models and the benchmark model GM-RV for each equity market index.

	$\sigma^2_{w1,w2}$				$\tau_{w1,w2}$			
RP metric	France	Germany	UK	US	France	Germany	UK	US
Panel A: flexible weighting scheme								
Naive	0.000	0.032	0.013	0.000	0.000	0.001	0.005	0.000
RC_c	0.073	0.661	0.662	0.213	0.095	0.022	0.492	0.067
$Hansen_c$	0.041	0.475	0.413	0.039	0.000	0.013	0.344	0.031
Panel B: restricted weighting scheme								
	σ^2_{w2}				τ_{w2}			
Naive	0.000	0.134	0.184	0.000	0.000	0.000	0.0167	0.046
RC_c	0.183	0.618	0.348	0.173	0.058	0.108	0.352	0.386
$Hansen_c$	0.097	0.574	0.217	0.073	0.051	0.083	0.276	0.282

Table 9 Multiple testing for out-of-sample MSE losses using alternate variance proxies

The table presents the p-values of the White's (2000) reality check (RC) test and the Hansen's (2005) superior predictive ability (SPA) test when we use variance proxies of GARCH (1, 1) i.e., $h_{i,t}$ and square of implied volatility indices i.e., IV^2 for respective markets replacing RV. The forecasting errors i.e., MSE of the competing models are compared against the GM-RV benchmark model using a fixed out-of-sample forecasting scheme. Parameters are estimated using for sample period January 1999- December 2011 and forecasting period is January 2012-December 2016. Panels A, B, C and D present results for France, Germany, the UK and the US, respectively. First, we present naïve p-values that are the bootstrap RC p-values for bi-model predictive ability test: bootstrapped p-values by comparing forecasting errors of the best model among $k = 26$ competing models relative to the benchmark model. Following Hansen (2005) we compute the test statistics for the consistent (centre, c) bound for both RC test and Hansen test. The test-statistic p-value is computed from 1000 bootstraps resamples and smoothing parameter q = 0.25. For clarity a p-value larger than 0.05 shows that no competing model outperforms the benchmark model i.e., null hypothesis cannot be rejected, whereas value <0.05 shows that at least there is one model that has superior predictive ability (lower MSE or forecasting errors) than the benchmark GM-RV model i.e., null is rejected. The $\sigma^2_{w1,w2}$ and the $\tau_{w1,w2}$ refers to the results of the forecasting errors of daily total and monthly long run volatilities when the weights for the beta polynomial for each input variable in the MIDAS filter are estimated, while σ^2_{w2} and $\tau_{,w2}$ refers to the model comparison results when $w_1 = 1$ and w_2 is estimated for the beta polynomial of each input variable in the MIDAS filter, respectively.

	$\sigma^2_{w1,w2}$		$\tau_{w1,w2}$		σ^2_{w2}		τ_{w2}	
RP metric	$h_{i,t}$	IV^2	$h_{i,t}$	IV^2	$h_{i,t}$	IV^2	$h_{i,t}$	IV^2
Panel A: France								
Naive	0.000	0.000	0.000	0.001	0.000	0.020	0.002	0.000
RC_c	0.167	0.139	0.361	0.398	0.319	0.283	0.531	0.346
$Hansen_c$	0.051	0.073	0.173	0.123	0.143	0.053	0.073	0.139
Panel B: Germany								
Naive	0.000	0.004	0.000	0.000	0.132	0.097	0.001	0.003
RCc	0.337	0.184	0.003	0.013	0.461	0.579	0.046	0.031
$Hansen_c$	0.064	0.089	0.017	<0.001	0.178	0.383	0.021	0.018
Panel C: the UK								
Naive	0.004	0.024	0.000	<0.001	0.161	0.107	0.000	0.005
RC_c	0.029	0.083	0.145	0.278	0.264	0.377	0.081	0.605
$Hansen_c$	0.009	0.037	0.042	0.269	0.178	0.226	0.038	0.38
Panel D: the USA								
Naive	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.001
RC_c	0.637	0.389	0.086	0.193	0.074	0.272	0.394	0.921
$Hansen_c$	0.000	0.023	0.009	0.041	0.053	0.134	0.287	0.531