Volatility Forecasting Performance at Multiple Horizons

For the degree of Master of Science in Financial Economics at Erasmus School of Economics, Erasmus University Rotterdam

Author: Sharon Vijn
Supervisor: Dr. R. (Rogier) Quaedvlieg
Second Assessor: Dr. Q. (Qinghao) Mao

August 24, 2017

ABSTRACT

This paper compares different well-known volatility models in terms of the in-sample and out-of-sample fit for different horizons to see which models perform best at several horizons. The return series of the S&P500 and the Mexican IPC are used to answer this question. The volatility is forecasted at the one day, one month, six months, one year and two year horizon under different distributions. Besides individual forecasts, forecast combinations are used as well to forecast volatility. All these forecasts are evaluated with the MSE and compared with the Diebold Mariano test and the Model Confidence Set. It can be concluded that in the short run there is not one model that outperforms other models. Half of all models seem to perform equally. Forecast combinations based on the trimmed mean and MSE ranks provide the most accurate forecasts in forecasting volatility in one or two years from now for the S&P500. Forecasting in the long-run for the IPC can be done most accurately by using GJR-GARCH.

Keywords: Forecasting, Multiple Horizons, Forecast Combinations, Model Confidence Set

ERASMUS UNIVERSITEIT ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS

CONTENT

1. INTRODUCTION

The return of almost every security is affected by fluctuations in the price. Therefore it is crucial to create tools to forecast these fluctuations called volatility for both financial institutions and researchers as well as for regulators. Volatility forecasting is one of the most important principles in risk management, asset allocation and option pricing. Volatility means a deviation from the mean which corresponds to risk. The more accurate the volatility forecast, the better one can determine the asset price which is very valuable.

Due to the excessive activity in forecasting volatility, various researchers developed a large amount of (sophisticated) models that try to explain the movements in financial asset volatility. This makes it interesting to test different models against simple historical models to see whether they outperform or not. Existing research on this topic is ambiguous due to, among other things, different time series, sample periods, distributions, loss functions, forecast horizons and what proxy to use for realized volatility. Poon and Granger summarized 93 papers on the forecasting performance of many models in 2003 and found that in almost 50% of the cases, the regression-based models were not able to outperform historical, naïve models. In almost all other cases, the asymmetric ARCH models particularly performed best.

A lot of research exists on forecasting volatility one day or one month ahead. Conclusions are based on forecasting in the short run but this does not mean the same is true for forecasting in the long run. Figlewski (2004) provided an extensive review on forecasting volatility in the long run and concludes that GARCH(1,1) is not good at forecasting in the long run but the simple historical methods actually are. More recently, Brownlees, Engle and Kelly (2012) showed that forecasting in the long run gets accompanied by more risk and therefore the loss functions will be higher.

Forecasting longer horizons is interesting and beneficial. Most money managers will agree on the fact that forecasting volatility one day in advance is insufficient. Also, it is very plausible that the best way to predict volatility in two years is very different from the method to forecast volatility in two weeks. So there is not one answer to 'what is the best forecasting method?' Also, the expansion in trading derivatives increased interest in forecasting in the long run since the lengthening of the written contracts. The aim of this paper is therefore to research what models provide most accurate forecasts in the short run and what models could be best used to forecast one or wo years ahead from now. The forecast horizon is extended to include one day, one month, six months, one year and two years ahead forecasts. To measure if the forecasting performance is statistically different among models, Diebold and Mariano's (1995) test for equal predictive accuracy will be applied. This test compares the loss functions of two models. The loss function used in this study is the MSE.

This paper also aims to conclude on whether the more complex models do provide more accurate forecasts than the simpler models. The models used in this study are the Simple and Exponential Moving Averages (SMA, EWMA), ARMA (Auto Regressive Moving Average), ARCH (Auto Regressive Conditional Heteroscedasticity), GARCH (Generalized ARCH), EGARCH (Exponential GARCH) and GJR-GARCH (Glosten-Jagannathan-Runkle GARCH). All ARCH models are tested under three different distributions as well: the Normal, Student's T and Generalized Error distribution. The return series of the S&P500 and the Mexican IPC from 01/03/2000 to 5/31/2017 are used to produce the results.

Another purpose is to answer the question whether to combine multiple forecasts of the same variable or to identify one single best forecasting model. To answer these questions, all models need to be compared and evaluated for each maturity to determine which forecast(s) are most accurate. Since the number of forecasts is quite large, the Model Confidence Set introduced by Hansen, Lunde and Nason (2011) offers a solution to this. A Model Confidence Set is a set of 'best' models for a given level of significance. The Model Confidence Set procedure is applied to all individual forecasts.

This paper is related to the extensive literature on volatility forecasting but it adds value in multiple ways. First of all not only individual forecasts at multiple horizons will be evaluated, forecast combinations will be examined as well to see if there is a difference between using combinations when forecasting in the short run or in the long run. Individual forecasts and forecast combinations are compared statistically, just like all other forecasts. Last but not least three different distributions are applied to see whether this improves accuracy.

The main conclusion of this study is that using different distributions than the normal distribution yield high MSEs. Models that perform well in the short run like ARMA or EWMA are not likely to be a guarantee that they also perform well in the long run. At the one day and one month horizon it is hard to draw conclusions. In the long run some clear results become visible. For the S&P500 the forecast combinations 'trimmed mean' and 'MSE ranks' provide the most accurate forecasts. In case of the more volatile IPC it is the GJR-GARCH that outperforms all other models and forecast combinations.

The remainder of this paper is structured as follows. The next section describes the theoretical framework, followed by a summary of the findings hitherto. Section 4 describes and analyses the data and section 5 explains the methods more closely and adds some literature about the methodology as well. Section 6 presents the results and the last section, section 7, summarizes and concludes. Thereafter the references and appendices can be found.

2. THEORETICAL FRAMEWORK

Before reviewing the existing literature on volatility forecasting, the existing models and some important concepts are discussed. This section starts with explaining realized volatility, the historical volatility models and the stylized facts. Thereafter the ARCH model and some of its extensions will be presented.

2.1 REALIZED VOLATILITY

First of all, it is necessary to define volatility. Volatility in financial markets can be explained as the spread of all likely outcomes of uncertain asset returns. In practice, volatility is generally calculated as the sample standard deviation, σ_t^2, which can be calculated as

$$\sigma_t^2 = \frac{1}{T-1}\sum_{t=1}^{T}(r_t - \mu)^2$$

where r_t denotes the return on day t and μ is the average return over the entire period T. Before high-frequency data became easily available, most researchers turned to the undesirable method of using daily squared returns as a proxy for daily volatility, assuming $\mu \approx 0$. This is however shown to be a very noisy estimator by, among others, Lopez (2001), Andersen and Bollerslev (1998) and Blair, Poon and Taylor (2001). The latter point out that the use of intraday 5-minute squared returns as a proxy increases accuracy up to three to four times. This sum of squared intra-period returns are called Realized Volatility (Poon, 2005).

The main advantage is that this proxy can be made readily accurate in a way that the interval over which the returns are calculated becomes negligibly small. This makes it possible to treat volatility as observable. More recent literature has concentrated on realized variance since high-frequency data has become widely and cheaply available. The additional information that intra-day data contains, makes it very attractive to use it (Andersen & Bollerslev, 1998). There are several other reasons why using realized variance (hereinafter RV) could be beneficial. For instance RV is non-parametric so there is no model risk. Also, RVs are simple to calculate. The only data one needs are market prices and most securities and instruments have this widely available. Finally, only information within the estimation interval is needed. So for instance in order to calculate the volatility for a period of 48 hours, everything needed are the intraday returns within those 48 hours.

2.2 SMA METHOD

A very simple method to forecast volatility is to calculate it from historical data. Historical volatility models (HIS) or naïve models are relatively easy to build and adjust. Conditional volatility is not modelled based on returns but directly on realized volatility which makes these models less restrictive and more prepared to respond to changes in volatility. The simplest HIS model is the random walk model. This model states that today's volatility predicts tomorrow's volatility:

$$\sigma_{t+1}^2 = \sigma_t^2$$

so only one variable is needed to predict tomorrow's volatility. The simple moving average (SMA) is based on this, however it uses older information as well

$$\sigma_{t+1}^2 = \frac{1}{\tau}(\sigma_t^2 + \sigma_{t-1}^2 + \cdots + \sigma_{t-\tau-1}^2)$$

where τ describes how many past observations will be used which makes this method very simple and with the improvement in intraday data, HIS models can provide very accurate forecasts.

2.3 EWMA METHOD

The exponentially weighted moving average (EWMA) is an expansion of the simple moving average by adding exponential weights so more weight is given to recent information and less to older

$$\sigma_{t+1}^2 = \sum_{i=1}^{\tau}(1-\lambda)^i \sigma_{t-i-1}^2$$

where λ is a constant and is sometimes called the 'smoothing constant'. Choosing a value for λ is an empirical issue but it is usually set to 0.94 following RiskMetrics approach. For both moving averages, the forecast is flat and volatility remains constant which means the h-day ahead forecast is the same as the one day ahead forecast.

2.4 ARMA

Besides the two moving averages above, there are autoregressive HIS models as well such as the simple regression method

$$\sigma_{t+1}^2 = \alpha + \beta_1 \sigma_t^2 + \beta_2 \sigma_{t-1}^2 + \cdots . + \beta_n \sigma_{t-n+1}^2$$

in which volatility depends linearly on its own previous values. If one adds past errors as well, this results in the Auto Regressive Moving Average (ARMA) designed by Peter Whittle in 1951

$$\sigma_t^2 = \alpha_0 + \sum_{j=1}^{p} \alpha_j \, \sigma_{t-j}^2 + \sum_{j=0}^{q} \beta_j \gamma_{t-j}$$

where γ_t are volatility errors that can be called white noise. ARMA is a linear Gaussian model and because there is a lot of supportive literature available on linear equations and Gaussian models, ARMA has been a commonly used model for a long period of time. Also, working with ARMA is quite simple and the model is found to be successful in analysing and forecasting data. Unfortunately it has its limitations as well. One of the stylized facts discussed in the next section, is that we usually see changes in volatility when looking at financial data. Because ARMA assumes constant volatility, this feature cannot be captured. Another shortcoming is that these models underperform when using data with strong asymmetry or data exhibiting strong cyclicality or time irreversibility (Knight & Satchell, 2007).

2.5 STYLIZED FACTS

Figure 1 and 2 plot the intraday 5-minute squared returns of the S&P500 and the IPC between 2000 and 2004. It becomes very clear that volatility indeed changes over time and that it is not just a constant plus some random noise. We do see quite a few peeks in the data. For instance around May 2000, the first quarter of 2001 and larger peeks around September 2001 – which is probably due to 9/11 – and the first quarter of 2002. The latter two are less visible for the IPC. From 2002 to 2004 volatility is very low with only one peak at the end of 2002. Appendix A shows figures of the realized volatility between 2004 and 2017 divided in periods of four years to see differences between the two indices.

Overall the IPC shows more short peaks in general, the peaks in the S&P500 take more time to disappear. The changes between high and low volatile periods are more subtle for the S&P500 compared to the sharp increases and decreases in IPC its volatility.

Figure 1: Realized Volatility S&P500 in-sample data

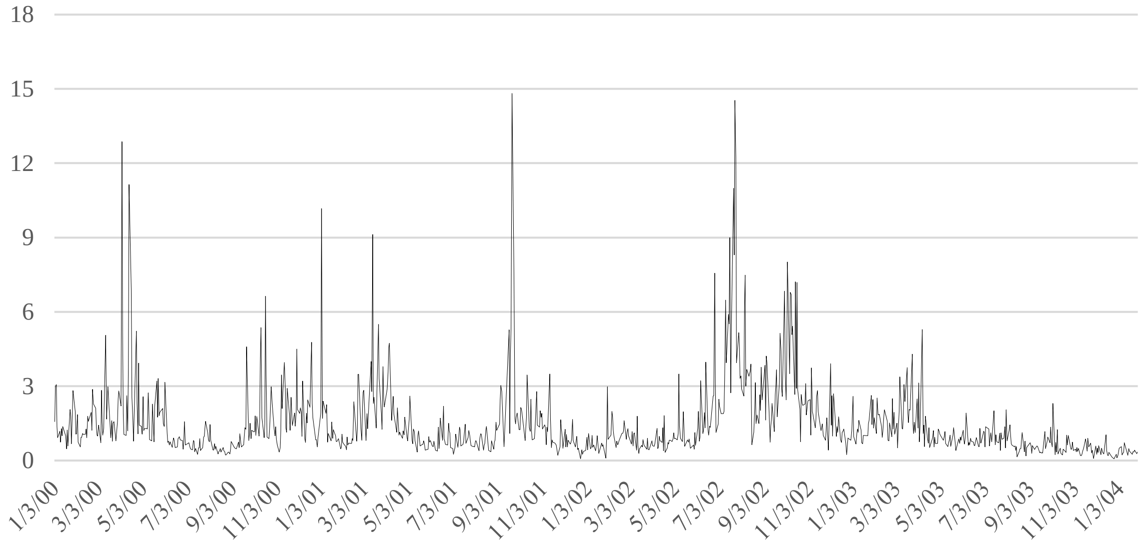
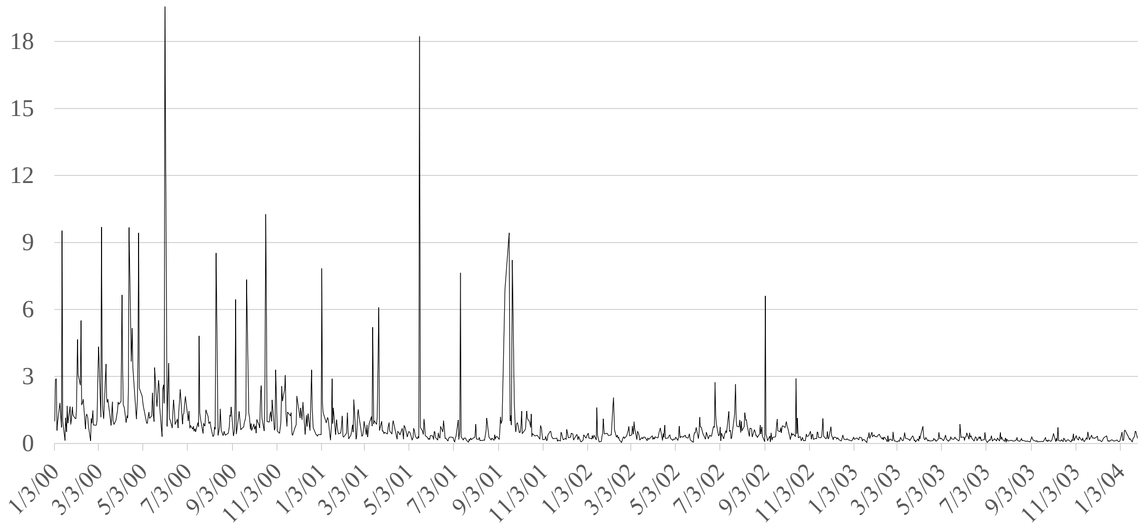


Figure 2: Realized Volatility IPC in-sample data



Aside from volatility being time-varying, financial data shows some other specific patterns which are called stylized facts. The following characteristics are usually observed when analysing financial data:

Volatility clustering – A phenomenon in financial time series is that low volatility is more likely to be followed by low volatility and that one turbulent trading day tends to be followed by another (Poon, 2005).

Leverage effect – Negative news leads to a fall in the stock price which shifts a firm's debt to equity ratio upwards. The firm has thus increased leverage i.e. higher risk. The corresponding stylized fact is that stock price volatility tend to increase more if the preceding day returns are negative relative to positive returns of the same magnitude (Christie, 1982). Not only the sign of the previous returns matters, the size does as well. Large negative and positive return shocks cause more volatility than small return shocks will (Engle & Ng, 1993).

Excess kurtosis and skewness – Most financial time series show excess kurtosis skewness. This leads to data that does not follow the normal distribution. Especially fatter left tails and higher peaks are well known features of financial asset returns. The normal distribution has a skewness of zero and a kurtosis of approximately three. Most financial time series are (far) above these values (Knight & Satchell, 2007).

Long memory – The autocorrelation of absolute or squared returns declines very slowly which means that volatility is highly persistent and that the effects of volatility shocks decay slowly. Poon (2005) shows that autocorrelation declines even slower for realized volatility. Figure 1 and 2 show that the long memory effect is more present in the returns of the S&P500 compared with those of the IPC.

Weak form market efficiency – Asset returns are usually not autocorrelated. If there exists some autocorrelation, it is only at lag one due to thin trading. In other words, returns are not predictable.

Co-movements in volatility – Returns and volatility across different markets or asset classes tend to move together. A shock in one currency can be matched with a shock in another currency. Or a shock in the stock market can be matched with a shock in the bond market. Especially correlation among volatility is strong and this effect is even bigger in bear markets or during financial crises (Poon, 2005).

Stylized facts are the cause of forecasting volatility being a difficult but interesting topic. It is the art to detect the time series properties and to use or create a volatility model that accounts for the stylized facts of financial market data.

2.6 ARCH(q) MODEL

In contrast to historical volatility models, the next models use asset returns as input instead of realized volatility. The Autoregressive Conditional Heteroscedasticity (ARCH) model is a more refined model that can be used in order to model volatility. ARCH(q) is designed by Engle in 1982 trying to capture volatility clustering, since this was a big shortcoming of ARMA. The model is, as the name says, Autoregressive because current volatility is related to previous period's volatility, this feature captures the volatility clustering aspect. Conditional to capture the time-varying aspect of volatility and Heteroscedastic to incorporate the autocorrelation often found in the squared residuals. Before explaining the model statistically, write returns as

$$r_t = \mu + \varepsilon_t$$
$$with\ \varepsilon_t = \sigma_t z_t$$

with $z_t \sim N(0,1)$. The ARCH(q) model than calculates conditional variance, σ_t^2, as

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \cdots + \alpha_q \varepsilon_{t-q}^2 \quad i.e.\ \ \omega + \sum_{j=1}^{q} \alpha_j \varepsilon_{t-j}^2$$

with q being the amount of lags and α being the ARCH parameters calculated through maximizing the likelihood of ε_t. Both ω and α_j must be equal to or larger than zero to establish a positive conditional variance. If $\sum_{j=1}^{q} \alpha_j < 0$, ARCH is stationary as well. Volatility is thus conditional on the squared residuals and because they differ in time, the model is time-varying. The formula above describes the one-step-ahead forecast. The multi-step-ahead forecast relies on the assumption that E[$\varepsilon_{t+\tau}^2$] = $\sigma_{t+\tau}^2$.

In theory, one could choose any value for q. This study sets q equal to one. σ_t^2 is actually a function of the information available at time t-1. In this case, the conditional variance only depends on one single observation: the past squared residual returns (Knight & Satchell, 2007). Despite the ability of the ARCH(q) model to capture volatility clustering, this model is not suited for variance effects that retain for a longer period of time. Trying to overcome this problem, Bollerslev and Taylor designed the Generalized ARCH (GARCH) model in 1986.

2.6.1 GARCH(p,q)

The difference between ARCH(q) and GARCH(p,q) is that the latter includes more dependencies and therefore it allows changes in volatility to occur more slowly. GARCH tries to capture another stylized fact: the long memory effect. To be specific, it includes lagged values of σ_t^2 as well, which results in the following identification

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_j \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$$

with α, β and ω being non-negative and α + β smaller than, but close to 1 in order for the model to be stationary. Note that for all ARCH models, σ_{t-j}^2 is not the same as the realized variance used in the HIS models. It is the forecasted volatility in the previous period. So tomorrow's volatility depends on today's volatility that was forecasted yesterday. The past squared residuals capture the high frequency effects and the lagged variance captures long term influences. So the expected volatility is a combination of the long run volatility and the expected volatility for the last few days. A big advantage of the GARCH model relative to EWMA for instance is that if today is a day of high volatility, the EWMA predicts all future days to be highly volatile as well whereas GARCH assumes that variance will move towards its average value in the long run. If q and p are both equal to one, we can write GARCH(1,1). The one-step ahead forecast of GARCH(1,1) is known at time t and will be

$$\sigma_{t+1}^2 = \omega + \alpha_1 \varepsilon_t^2 + \beta_1 \sigma_t^2$$

The two-step-ahead forecasts can be calculated by assuming $\mathrm{E}[\varepsilon_{t+1}^2] = \sigma_{t+1}^2$

$$\sigma_{t+2}^2 = \omega + \alpha_1 \varepsilon_{t+1}^2 + \beta_1 \sigma_{t+1}^2 = \omega + (\alpha_1 + \beta_1)\sigma_{t+1}^2$$

Similarly,

$$\sigma_{t+3}^2 = \omega + (\alpha_1 + \beta_1)\sigma_{t+2}^2$$

and so on until eventually the long-horizon forecast i.e. two years ahead will be the long-run average variance (Christoffersen, 2012). GARCH is simple and able to capture time variation and the long memory effect. One of the limitations however is that it can be difficult to fit the data, especially when more than one lag is used. Another shortcoming is that this model does not take asymmetries into account. GARCH might forecast volatility too low after a large shock in the asset price and too high in the case of a positive return shock. Finally, the non-negativity constraints on α, β and ω can create difficulties in estimating models. The next model, EGARCH, offers a solution to the latter problems.

2.6.2 EGARCH(p,q)

In 1991 Nelson presents the Exponential GARCH (EGARCH) model where the above mentioned constraints are not necessary because conditional variance is specified in logarithmic form

$$\ln(\sigma_t^2) = (1-\alpha_1)\alpha_0 + \alpha_1 \ln(\sigma_{t-1}^2) + \theta\left(\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right) + \gamma\ [\frac{|\varepsilon_{t-1}|}{\sigma_{t-1}}]$$

where β, γ and α are constants without constraints. θ is typically negative, so positive return shocks have less impact on volatility than negative shocks will have. γ captures the size effects because it depends on the absolute residual values. Larger shocks have a bigger influence on volatility than small shocks. Note that standard deviation is used as an input to calculate conditional variance. A reason for this could be that variance is less stable in computer estimation and standard deviation has the same unit as the mean instead of its square (Poon, 2005). Tsay (2002) illustrates how to define the one-step-ahead forecasts when the innovations are standard Gaussian, by taking exponentials

$$\sigma_{t+1}^2 = \sigma_t^{2\alpha_1} \exp[(1-\alpha_1)\alpha_0] \exp[\theta\left(\frac{\varepsilon_t}{\sigma_t}\right) + \gamma\left[\frac{|\varepsilon_t|}{\sigma_t}\right]]$$

the τ-step-ahead forecasts are defined as

$$\sigma_{t+\tau}^2 = \sigma_t^{2\alpha_1}(\tau-1)\exp[(1-\alpha_1)\alpha_0] * \{\exp[0.5(\theta+\gamma)^2]\,\Phi(\theta+\gamma) + \exp[0.5(\theta-\gamma)^2]\,\Phi(\theta-\gamma)\}$$

where Φ is the cumulative density function of the standard normal distribution.

2.6.3 GJR-GARCH

Another model that takes the asymmetry effect into account is the GJR-GARCH model designed by Glosten, Jagannathan and Runkle in 1993 where conditional variance is estimated as

$$\sigma_{\mathrm{t}}^2 = \omega + \sum_{j=1}^{q} (\alpha_j \varepsilon_{t-j}^2 + \delta_j D_{j,t-1} \varepsilon_{t-j}^2) + \sum_{j=1}^{p} \beta_i \sigma_{t-i}^2$$

with δ_j being the leverage term and D is a dummy variable which takes the value 1 if ε_{t-1} < 0 and 0 if otherwise. In this model α, β and ω must be non-negative and α + β must be smaller than one, but again still close to one for stationarity. An additional restriction is that γ should be equal to or larger than zero. For the one-step-ahead forecast the equation becomes

$$\sigma_{t+1}^2 = \omega + \beta_1 \sigma_t^2 + \alpha_1 \varepsilon_t^2 + \delta_1 \varepsilon_t^2 D_t$$

and the multi-step-ahead forecast is

$$\sigma_{t+\tau}^2 = \omega + (0.5(\alpha_1 + \gamma_1) + \beta_1) \sigma_{t+\tau-1}^2$$

The GJR-GARCH model takes the asymmetric effect into account by adding the leverage term. The forecasted volatility will be higher when there was a loss instead of a positive return. Volatility persistence can change quite fast when the return changes sign.

3. LITERATURE REVIEW

After discussing all the models and the stylized facts often found in financial time series data, this section summarizes some findings thus far. Since modelling and forecasting volatility is, and has for several decades been, a very attractive topic for researchers a lot of different outcomes have been published by a large volume of experts. Findings are ambiguous due to several reasons. Poon and Granger (2003) provide an extensive review of 93 published and working papers that study the forecasting performance of a broad range of models. They find that GARCH models outperform ARCH models but asymmetric models perform even better. They also show that the simple historical volatility models are able to outperform the more complex regression based models in almost half of the cases. Other researchers that prefer HIS models over ARCH models are Taylor (1986, 1987), Figlewski (1997), Figlewski and Green (1999), Andersen, Bollerslev, Diebold and Labys (2001) and Taylor (2004). The main conclusion they all draw is that when there is a change in the volatility level, parameter estimation gets unstable and the predictive power suffers.

The ARCH models however, have a lot of proponents as well. Akgiray (1989) was one of the first researches who tested the predictive power of ARCH models and finds that GARCH outperforms EWMA and SMA in all different periods and under all sorts of evaluation measures. Figlewski (1997) agrees on this but only when forecasting over a short horizon. If ARCH models outperform HIS models, it is usually the conclusion that asymmetric models perform best. Brownlees, Engle and Kelly (2012) find that asymmetric models, especially GJR-GARCH perform well across assets. Hansen and Lunde (2005) compared 330 ARCH-type models and find no evidence that GARCH(1,1) is outperformed when forecasting exchange rate volatility, but models that incorporate the leverage effect such as GJR-GARCH or EGARCH are preferred when analyzing stock return volatility. Differences between GJR-GARCH and EGARCH seem inconclusive. For instance Pagan & Schwert (1990) and Cao and Tsay (1992) favor the EGARCH model, while Brailsford and Faff (1996) and Taylor (2004) prefer GJR-GARCH. Studies that find no pronounced results are most often studies that use squared daily returns to proxy actual volatility. Due to the noise in this proxy, the (small) differences between models become indiscernable (Poon, 2005).

A lot of papers focus on short-term forecasting like one-day or one-week ahead forecasts. Also, the most widely used risk measures Value-at-Risk (VaR) and Expected Shortfall

(ES) focus on short-term risks while they are often misused in measuring long-term risks. Figlewski (2004) is one the researchers who focuses on predicting long horizon volatility. He examines the performance of GARCH(1,1) and finds difficulty in forecasting volatility over long horizons with this model. When forecasting with GARCH(1,1) for more than one period ahead, the forecasts do not incorporate new information about future shocks. It will just converge to the long run variance at a rate determined by $\alpha_1 + \beta_1$. Figlewski (1997) showed that forecasts from simple historical methods are more accurate at horizons longer than six months than model-based forecasts. Alford and Boatsman (1995) agree on this. Figlewski (1997) concludes that forecast accuracy is higher for longer horizons than for shorter horizons because it seems to be true that today's variance will move towards its long run variance in a couple of years from now.

Brownlees et al. (2012) do not agree with this and say forecasts deviate more from reality because there is always an extra type of risk that the risk itself will change. They also state that asymmetric models provide more accurate one-day and one-week ahead forecasts. At the one-month horizon the difference between asymmetric and symmetric models becomes less visible because recent negative news has a lower influence on predicting volatility a few weeks ahead. They also do not deny the presence of fat tails in financial time-series data but they do not find benefits to use a Student's t-distribution instead of a normal distribution. Franses and Ghijsels (1999) even find that the performance of the GARCH model under the Student's t-distribution performs a lot worse than using the normal distribution in terms of out-of-sample performance. Hansen and Lunde (2005) come up with the same conclusion for IBM stock return data. Wilhelmsson (2006) studies the performance of the GARCH model under nine different error distributions. He shows that the chosen loss function can have a lot of impact on the results and concludes that using a leptokurtic but symmetric distribution i.e. the Student's t-distribution, improves results substantially. The Mean Absolute Error (MAE) and the Heteroscedasticity-adjusted MAE (HMAE) are used as loss functions to evaluate the use of different distributions because, according to him, the MSE criterion is sensitive to large return shocks.

Besides comparing individual forecasts, this study discusses forecast combinations as well. Forecast combining, or sometimes called forecast averaging, is a method to combine different forecasts into one forecast. Many studies have shown that combinations of forecasts have lower loss functions than the one best individual model. Makridakis and

Winkler (1983) were one of the first who find large gains from averaging the forecasts with simple methods and more recently Stock and Watson (2001) agree on this as well. They find that especially the average or median forecast and forecasts weighted based on the inverse MSE perform very well. They add that forecast combinations have a superior performance at the one, six and twelve month horizons and that it is best to combine as many forecasts as possible.

One explanation of why forecast combinations might work has to do with the difference in degrees of adaptability. One model may adapt quickly where another model adjusts very slowly. The combination of this probably works better than one model in isolation. The second possible explanation has to do with the misspecification bias. It is quite dubious to believe that the same model outperforms all other models at all times. It can be expected that the best performing model changes over time. Combining forecasts can create a more robust forecast, protected against such misspecification. Another somewhat similar argument is that the risk of choosing the wrong method can be very serious. When averaging forecasts, the choice of the methods become less important. The outcome does not depend on one model anymore (Makridakis & Winkler, 1983).

Of course besides reasons why one should combine forecasts, there are also arguments against using forecast combinations. Estimation errors that harm combination weights is one of the main problems for many combination techniques. Also, non-stationarity in the underlying is one of the reasons to combine forecasts but on the other hand this phenomenon creates unstable combination weights as well. Therefore it can be very hard to find a set of weights that perform well.

Empirical findings are different among studies but there are some general conclusions that can be drawn. Most researchers suggest that simple combination schemes are actually better than more complex weighting schemes. Examples of simple combinations are the arithmetic average or weights based on the inverse MSE. Combinations based on in-sample performance usually lead to poor predictive ability. Simple combinations are combinations that do not require estimating (many) parameters since the weights are already known. This is exactly the reason why they are preferred over more complex combinations. If the weights need to be estimated, parameter estimation errors are likely to arise (Timmermann, 2006).

4. DATA

The data used is gathered from the Oxford-Man Institute's Realized Library which contains daily (close to close) financial returns and daily measures of how volatile financial assets or indexes were in the past. Data is originally from the Reuters DataScope Tick History database. Realized measures ignore the variation of prices overnight and sometimes the variation in the first few minutes of the trading day when recorded prices may contain large errors. In the realized library, data is available from 01-03-2000 up until today. The S&P500 and the Mexican IPC will be used as equity indices. The S&P500 represents 500 large-cap companies traded on American stock exchanges. The IPC is an index of 35 companies that trade on the Mexican Stock Exchange. This provides some interesting insights in the differences between an index representing a developed country and one that represents a developing country. Since they have approximately the same number of transactions in the Realized Library, a fair comparison can be made.

An important consideration is the length of the in-sample data. Either a longer sample which implies more precise estimates but probably structural breaks are included, or a short sample which is less precise but there is less risk of estimating across a structural break. Alford and Boatsman (1995), Figlewski (1997) and Figlewski and Green (1999) all agree on the importance of having at least a long enough estimation period to make accurate volatility forecasts. But maybe instead of using the same in-sample data for all forecasting horizons, it might be better to use shorter samples in trying to forecast volatility over the next day or month and longer samples when trying to predict volatility in one or two years from now. Figlewski (2004) finds that using long historical samples (i.e. 4 to 5 years of data) turned out to be the most accurate in all cases. Therefore following Figlewski (2004) and Christoffersen (2012) 1,000 daily observations i.e. approximately 4 years (from 01-03-2000 to 01-26-2004) are used for the in-sample data which is said to be a fairly good general rule of thumb. The out-of-sample forecasted period is 01-27-2004 to 05-31-2017. This period covers both calm and stormy periods.

4.1 DESCRIPTIVE STATISTICS

Table 1 presents the descriptive statistics of the daily returns series from 01-03-2000 to 05-31-2017 obtained from the Realized Library. In total there are 4,351 observations of which 1,000 in sample and 3,351 out of sample for both indices. The table shows that the S&P 500 and IPC are clearly not normally distributed. The Jarque-Bera test is used to test whether a sample follows a normal distribution. The JB test statistic is respectively 11,266 and 4,768 with a p-value of 0.000 which means the null hypothesis can be rejected at any level of significance. The excess kurtosis is definitely higher than three for both indices as well , which means there are more chances of extreme outcomes compared to a normal distribution.

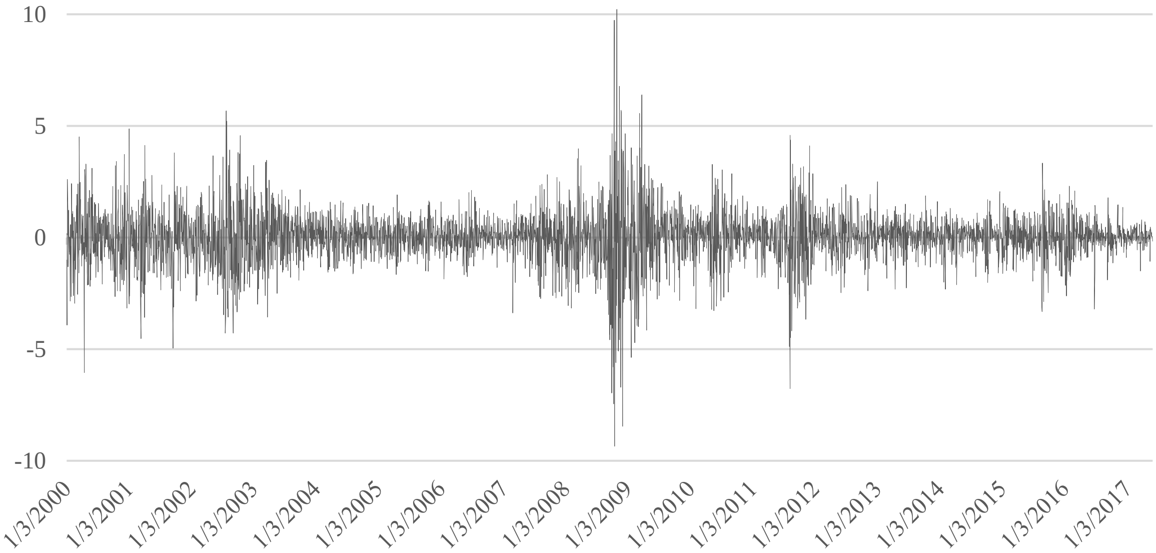
Table 1: Descriptive statistics of the daily return series from 01-03-2000 to 05-31-2017.

	Daily Average	Maximum	Minimum	Daily Variance	Skewness	Kurtosis	JB statistic	JB p-value
S&P 500	0.010	10.220	-9.351	1.379	-0.171	10.882	11,285	0.000
IPC	0.038	9.953	-8.261	1.676	-0.003	8.128	4,768	0.000

Statistic are reported in percentages except the JB statistic and its p-value. Daily average and daily variance are both unconditional. In total there are 4,351 observations for both indices. Outliers are not removed from the dataset.

Figure 2 plots the daily returns of the S&P 500 for the entire period. What can be seen immediately is that relatively calm periods are followed by more stormy periods which is one of the stylized facts discussed before. Around 2009 and 2011/2012 we see a very turbulent period as well as from 2000 up until 2004. This is not very different for the IPC.

Figure 2: Daily returns S&P500 from 01/03/2000 to 5/31/2017



4.2 STATISTICAL TESTS

Analysing data and testing for stylized facts are important first steps in determining which model forecasts best since the out-of-sample forecast performance might be influenced by the in-sample fit. Besides testing for normality it is convenient to test for ARCH effects i.e. whether the data is non-linear. Next it is useful to test whether the leverage effect is present in the time-series or not with the sign bias test by Engle and Ng (1993) which demonstrates whether the residuals in the GARCH-model are sign biased or not. Finally the Augmented Dickey-Fuller test for unit root (1979) is used to test if a time-series is stationary.

4.2.1 ENGLE'S ARCH LM TEST

Engle's ARCH test is a Lagrange multiplier test to determine the significance of ARCH effects by running a regression of the squared residuals on lagged squared residuals and a constant

$$\varepsilon_t^2 = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \cdots + \alpha_j\varepsilon_{t-j}^2$$

with the null suggesting that there is no autocorrelation in the squared residuals:

$$H_0 = \alpha_0 = \alpha_1 = \cdots = \alpha_j = 0$$

This can be done on the residuals of an ARMA(1,1) model or an ARCH model. Table 2 presents the results of the ARCH LM test at lag five for both series. The results are based on a regression of the residuals on an ARCH test, but was performed on the residuals of an ARMA(1,1) model as well but the outcome did not change neither do they change if lags differ.

Table 2: Engle's ARCH LM test results at lag 5

	Engle's LM Test Statistic	P-value
S&P 500	67.248	0.000
IPC	68.493	0.000

The in-sample data is fitted to an ARCH model and this table presents regression results with the squared residuals as dependent variable and the lagged squared residuals (up until the fifth lag) as independent variables.

For both series the null gets rejected which means there is autocorrelation in the squared residuals. In other words this means that there is conditional heteroscedasticity in both time series. This suggests that models which do not assume constant variance could provide more accurate forecasts.

4.2.2 SIGN AND SIZE BIAS TEST BY ENGLE AND NG

A sign bias test can be performed to test whether positive and negative shocks have a different impact on volatility. A more extensive test involves testing if volatility depends on both the size and sign of shocks introduced by Engle and Ng in 1993. The regression looks as follows

$$\hat{\varepsilon}_t^2 = \alpha_0 + \alpha_1 S_{t-1}^- + \alpha_2 S_{t-1}^- \hat{\varepsilon}_{t-1} + \alpha_3 S_{t-1}^+ \hat{\varepsilon}_{t-1} \; with \, S_{t-1}^+ = 1 - S_{t-1}^-$$

with the dummy variable, S_{t-1}^-, indicating the sign bias that takes the value of one if the past residual is negative and zero if it is positive. $S_{t-1}^- \hat{\varepsilon}_{t-1}$ indicates the negative size bias and $S_{t-1}^+ \hat{\varepsilon}_{t-1}$ the positive size bias. The null, H$_0$: $\alpha_1 = \alpha_2 = \alpha_3 = 0$, suggess there is no asymmetry at all in the residuals. A significant α_1 suggests sign bias and a significant α_2 and α_3 suggest negative size bias and positive size bias respectively. The sign bias test examines if positive and negative shocks affect future volatility in a different way. Literature points out that negative returns have a larger influence on volatility than positive returns of the same magnitude. The negative size bias tests whether large and small negative shocks have a different impact on future volatility and the positive size bias does the same except than for positive shocks.

The results are reported in table 3. The S&P500 shows no significant sign bias or positive size bias but it does indicate negative size bias which means large negative shocks have a larger influence on future volatility than small shocks. The IPC shows both significant negative and positive size bias and no significant sign bias either, thus both large negative and positive shocks have a larger influence on volatility than small shocks. Both indices do not show evidence that negative returns influence volatility more than positive returns. The null however can be rejected at the 1% significance level for both series which gives reason to assume there is a leverage effect. This gives reason to use asymmetric models.

Table 3: Engle and Ng's Sign and Bias test results.

	Sign Bias	Negative Size Bias	Positive Size Bias	Joint Effect
S&P 500	-1.039(0.299)	-4.94(0.00***)	0.598(0.550)	9.59(0.00***)
IPC	1.499(0.134)	-3.76(0.00***)	3.12(0.0***)	10.19(0.00***)

Sign and Bias test is based on fitting a symmetric GARCH(1,1) model to the in-sample data ranging from 01/03/2000 to 01/26/2004. The obtained squared residuals are used as dependent variable in the regression. T-values are shown with p-values in brackets. ***corresponds to a significance level of 1%.

4.2.3 AUGMENTED DICKEY-FULLER TEST

Finally the Augmented Dickey-Fuller (ADF) test for stationarity is important to choose a suitable model. The regression underlying the test looks as follows

$$y_t = \alpha + \delta t + \phi y_{t-1} + \beta_1 \Delta y_{t-1} + \cdots + \beta_p \Delta y_{t-p} + \varepsilon_t$$

where y_{t-1} is the absolute (lagged) return and p the amount of lags. The null of a unit root is H$_0$: $\phi = 1$ and the alternative hypothesis is $\phi < 1$. The ADF tests the in-sample absolute returns and accepting the null means the series is non-stationary and thus assume it follows a random walk. The DF statistic reported below is calculated as

$$DF = \frac{\widehat{\emptyset} - 1}{SE(\widehat{\emptyset})}$$

Table 4 shows that for both the S&P500 and the IPC the P-value is significant at the 1% level so the null can be rejected suggesting the time series do not follow a random walk but they are stationary.

Table 4: Augmented Dickey-Fuller test results

	Statistic	P-value
S&P 500	-15.593	0.001
IPC	-15.697	0.001

The in-sample absolute returns are tested. In-sample data ranges from 01/03/2000 to 1/26/2004.

The theory behind the ARMA model is based on stationary time series, so therefore it is especially important to consider this feature when applying an ARMA model. Since both series are stationary, the ARMA model can be used to forecast volatility.

After analysing the data, one can conclude that both time series are not normally distributed, they are stationary, there is some conditional heteroscedasticity in the data and the leverage effect is present. The following section fits the data to the models and describes the methods used to compare the in-sample fit and the out-of-sample forecasts.

5. METHODOLOGY

What can be concluded from the literature review and the data analysis is that volatility is time-varying and predictable. After a thorough discussion of the several models and analysing the data, the next step is to estimate the parameters in the models. The difficulty with estimating the autoregressive models is that the conditional variance should be estimated together with the parameters of the model. The method used to find the parameters is the maximum likelihood estimation. This method maximizes the most likely parameters through an iterative procedure given a log-likelihood function. The estimated parameters can be found in appendix B. When the model fitting is completed, the goodness of fit can be compared. In other words, how well does each of the models fit the in-sample data. The comparison can be found in section 5.1. Afterwards the parameters are then used in order to forecast volatility using a rolling window which will be discussed in the section 5.2.

5.1 IN-SAMPLE MODEL FITTING AND EVALUATION

A general method to evaluate the model fitting is to use an information criteria. A very well-known critera is the Akaike Information Criteria (1973) which is defined by

$$AIC = -2logL(\theta) + 2k$$

where $logL(\theta)$ is the maximized log-likelihood function and k is the number of parameters. Table 5 shows the AIC for all the autoregressive models. The smaller the value of AIC, the better the model fits the in-sample data. Of all models, EGARCH under the Student's T distribution has the lowest AIC so the in-sample data fits this model at best. ARMA(1,1) provides the worst in-sample fit. For all models, the non-normal distribution provides a better fit than the normal distribution. Something that was already expected from the data analysis before.

Table 5: Akaike Info Criterion (AIC) of all ARCH models.

	ARCH(1)	ARCH(1)	ARCH(1)	GARCH(1,1)	GARCH(1,1)	GARCH(1,1)
Model	Normal	Student's T	GED	Normal	Student's T	GED
S&P500	3.441	3.412	3.411	3.328	3.317	3.320
IPC	3.648	3.564	3.565	3.493	3.457	3.460
	EGARCH(1,1)	EGARCH(1,1)	EGARCH(1,1)	GJR-GARCH(1,1)	GJR-GARCH(1,1)	GJR-GARCH(1,1)
Model	Normal	Student's T	GED	Normal	Student's T	GED
S&P500	3.264	**3.261**	3.264	3.280	3.274	3.278
IPC	3.457	**3.438**	**3.438**	3.465	3.442	3.444

In-sample return data of S&P 500 and IPC ranging from 01-03-2000 to 01-27-2004 is fitted to all the ARCH models and the average AIC is reported. ARMA(1,1) is excluded because this model is fitted to realized volatility instead of returns.

5.2 FORECASTING PROCEDURE

The volatility in the out-of-sample period is forecasted through the in-sample data using a rolling window with a fixed number of observations. For example when forecasting on a daily basis, the first forecasted day uses the entire in-sample data. For the next day, the oldest day of the in-sample data is excluded and the first realized daily value is used in-sample to produce the next forecast. This is more accurate than using just a growing window because it ignores information from the distant past and the calculations are manageable since the number of observations stays the same. This procedure is repeated throughout the whole out-of-sample period.

Lastly, all models are analysed under three distribution assumptions: the Normal-, Student's t-, and General Error- distribution to see whether this provides more proper forecasts. h denotes the forecast horizon, so for each model the 1-step to h-step ahead forecast is computed. If h = 1, it means the 1-day ahead volatility is forecasted. If h = 21, the 1-month ahead forecasts are forecasted since there are approximately 21 trading days each month. This paper predicts volatility one day, one month, six months, one year and two years ahead which results in a daily volatility forecast path $\{\sigma_{t+h|t}^{(m)}\}$ where m denotes the model used. This method produces an array of overlapping forecast paths, with each path drafted from different conditioning information.

5.3 FORECAST EVALUATION

The volatility obtained must be evaluated to see how accurate the estimates are. Therefore the forecast errors must be calculated for each of the 1- to h-step ahead forecasts i.e. the difference between the forecasted volatility and the 'actual' volatility. If all forecasts and corresponding forecast errors for all models have been calculated, a loss function is necessary to assess all these forecasts. The model that yields a smaller average loss is more accurate and thus favoured. This sounds easy, but the difficulty is that an ex post proxy of volatility is needed as actual volatility.

5.3.1 MSE AND DIEBOLD MARIANO TEST

Section 2 presented the use of 5-minute intraday squared returns or Realized Volatility (RV). RV is an estimate of the true out-of-sample volatility used in this paper as a proxy of the actual volatility.. Knight and Satchell (2007) define this as the sum of squared intra-period returns from a Gaussian diffusion process.

This paper uses RVs which are calculated using five-minute returns. Using a shorter interval creates market microstructure problems i.e. noise in the data due to bid-ask spreads, non-trading and serial correlation (Figlewski, 1997). Liu, Patton and Sheppard (2015) studied the accuracy of almost 400 realized measures and find little to no evidence that the 5-minute RV is outperformed by any of the other measures.

To evaluate the forecasts, a large amount of statistical criteria is available. One of the most popular loss functions is the Mean Squared Error (MSE) which is defined by

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(\hat{\sigma}_t - \sigma_t)^2$$

which results in an average for the deviation of the estimated and realized standard deviation $(\hat{\sigma}_t - \sigma_t)$ and can be compared between the different models. Squaring the error gives larger weight to greater errors.

The model that performs best is the model with the lowest value for the MSE but without tests of significance, one cannot draw a conclusion. To overcome this problem, the Diebold Mariano (1995) (DM) test can be applied to determine which one of the two forecasts is significantly better. The DM test takes the difference between two loss functions resulting in a series d_{ij} with average $\bar{d}$. This average is zero if there is no difference between the forecasts which is the null hypothesis. The DM statistic looks as follows using a standard t-test

$$DM = \frac{\bar{d}}{\sqrt{\widehat{Var}(\bar{d})}}$$

For h-step ahead forecasts, the DM statistic must take autocorrelation into account because multi-period forecast errors are very likely to show this. Using Semin Ibisevic's Toolbox in Matlab the DM statistic is retrieved, taking into account autocorrelation and the sample variance is estimated using a Newey-West type estimator since they are robust to both heteroscedasticity and autocorrelation. Another option is to regress d_t on a constant and check whether this constant is significant or not, again with Newey-West standard errors. Both the MATLAB Tool and the regression are used to compare loss functions.

5.3.2 THE MODEL CONFIDENCE SET

Since this paper studies seven models in total under different distributions, it is more convenient to use the Model Confidence Set introduced by Hansen, Lunde and Nason (2011). A model confidence set (MCS) can be seen as a serial Diebold-Mariano test and is a set of best models, $\mathcal{M}^*$, out of the whole set of competing models, $\mathcal{M}^0$, for a given level of significance α. Most of the time there is not a single model that dominates all other models and therefore it can be favourable to identify a set of best models instead of just one. The forecasts are evaluated based on their relative performance to the other forecasts by means of the squared forecast error. The difference between two loss functions i and j is then called d_{ij}, just like the Diebold Mariano test does. Likewise, the assumption $\mu_{ij} = E(d_{ij,t})$ is made. The lower the MSE the better, thus model i is preferred to model j if $\mu_{ij} < 0$. The set of superior models is defined by

$$\mathcal{M}^* = \{i \in \mathcal{M}^0 : \mu_{ij} \leq 0 \text{ for all j} \in \mathcal{M}^0\}$$

The MCS procedure is based on two tests. First, an equivalence test, $\delta_{\mathcal{M}}$, which tests the following hypothesis

$$H_0 : \mu_{ij} = 0 \text{ for all i, j } \in \mathcal{M}$$

where $\mathcal{M} \subseteq \mathcal{M}^0$. The null suggests that the models in the set perform equally 'well' and is based on t-statistics. Second, an elimination rule, $e_{\mathcal{M}}$, which eliminates models from the set if $\delta_{\mathcal{M}}$is rejected. This procedure continues until $\delta_{\mathcal{M}} = 0$ and thus accepted. All the residual 'surviving' models in that set perform equally. The MCS algorithm works as follows:

1. Originally set $\mathcal{M} = \mathcal{M}^0$.
2. Test the null hypothesis using $\delta_{\mathcal{M}}$ at the level of significance α.
3. If H_0 is accepted, define $\mathcal{M}^* = \mathcal{M}$. Otherwise use the elimination rule $e_{\mathcal{M}}$to eliminate a model from $\mathcal{M}$ and repeat step 2.

The MCS procedure produces p-values as well for each of the models. The MCS p-value for model $e_{\mathcal{M}_j} \in \mathcal{M}^0$ is defined by $\hat{p}_{e_{\mathcal{M}_j}} = max_{i \leq j} P_{H_{0,\mathcal{M}_i}}$. With $P_{H_{0,\mathcal{M}_i}}$ being the p-value associated with the null hypothesis $H_{0,\mathcal{M}_i}$. If$\widehat{p}_i \geq \alpha$, the model will be included in $\mathcal{M}^*$.

Models with small p-values are thus likely to be not included in the set of best models and if they are, they are probably not one of the best alternatives. The p-value should not be interpreted as if some particular model is the best model. Hansen, Lunde and Nason (2011) provide a table to show how MCS p-values are calculated:

Table 6: Computation of MCS p-values.

p-value for $H_{0,\mathcal{M}_\kappa}$	MCS p-value
$p_{H_{0,\mathcal{M}_1}} = 0.01$	$p_{e_{\mathcal{M}_1}} = 0.01$
$p_{H_{0,\mathcal{M}_2}} = 0.04$	$p_{e_{\mathcal{M}_2}} = 0.04$
$p_{H_{0,\mathcal{M}_3}} = 0.02$	$p_{e_{\mathcal{M}_3}} = 0.04$
$p_{H_{0,\mathcal{M}_4}} = 0.03$	$p_{e_{\mathcal{M}_4}} = 0.04$
$p_{H_{0,\mathcal{M}_5}} = 0.07$	$p_{e_{\mathcal{M}_5}} = 0.07$
$p_{H_{0,\mathcal{M}_6}} = 0.04$	$p_{e_{\mathcal{M}_6}} = 0.07$
$p_{H_{0,\mathcal{M}_7}} = 0.11$	$p_{e_{\mathcal{M}_7}} = 0.11$
$p_{H_{0,\mathcal{M}_8}} = 0.25$	$p_{e_{\mathcal{M}_8}} = 0.25$
...	...
$p_{H_{0,\mathcal{M}_0}} = 1.00$	$p_{e_{\mathcal{M}_0}} = 1.00$

This table is copied from 'The Model Confidence Set' written by Peter R. Hansen, Asger Lunde and James M. Nason in 2011. Source: Econometrica, Vol. 79(2) pp. 453-497. This table is reported on page 462 in their paper.

One can see that some p-values for the null hypotheses do not coincide with the MCS p-values. For example the MCS p-value for $e_{\mathcal{M}_3}$, which is the third model that must be eliminated from the set, is larger than the p-value for $H_{0,\mathcal{M}_3}$. This is because the p-value of the null hypothesis tested before, $H_{0,\mathcal{M}_3}$, is already larger. If α would be five percent in this case, the first four models would be excluded from $\mathcal{M}^*$.

An estimate of the variance of d_{ij} must be made using a bootstrap. This paper uses a block-bootstrap procedure of 10,000 replications and a block length l long enough to capture the autocorrelation, if any, in the loss functions. The longer the horizon, the longer the block length should be due to losses at longer horizons tend to persist longer (Quaedvlieg, 2017). In this case, this means block lengths equal the forecast horizon. For instance when forecasting volatility one month ahead, the block length is 21. Using longer block lengths does not change the outcomes. The model confidence sets are all constructed with a confidence level of 95% which corresponds to an α of 0.05.

5.4 FORECAST COMBINATIONS

After identifying a single dominant forecast it is also interesting to create a combination of forecasts because forecast combinations gain from diversification benefits. The challenging part of this is the determination of the combination weights. Suppose some forecast, $\hat{\sigma}^2_{t+1,1}$ is significantly better than another forecast, $\hat{\sigma}^2_{t+1,2}$. So the expected loss is lower under , $\hat{\sigma}^2_{t+1,1}$ than under $\hat{\sigma}^2_{t+1,2}$. This means no one would choose $\hat{\sigma}^2_{t+1,2}$ over $\hat{\sigma}^2_{t+1,1}$ in isolation but a combination of those two forecasts can generate a smaller expected loss than just $\hat{\sigma}^2_{t+1,1}$. Apparently it seems that when $\sigma_{lossfunction1} > \sigma_{lossfunction2}$ and the correlation between the two loss functions is not equal to $\frac{\sigma_{lossfunction2}}{\sigma_{lossfunction1}}$, it would be optimal to combine those two forecasts (Timmermann, 2006). As stated in section 3, it has often been found that the sophisticated combinations are dominated by the more simple methods, therefore this paper only contains simple weighting schemes. The methods used in this study will be discussed briefly, starting with the simple mean.

The simple mean method calculates the arithmetic average of the forecasts at each point in time. In other words, the forecasts are equally weighted. The trimmed mean is like the simple mean with one difference. The forecasts are ordered and at each observation the highest ν% and the lowest ν% of the forecast values are removed before calculating the mean. The selection of forecasts that need to be removed is recalculated at each observation, so the weights are time-varying. This study uses different levels of trimming, depending on what is optimal for each model. Another similar method is the simple median method which calculates the median of the forecasts at each point in time.

A somewhat different method is the least squares weighting method. Using this method, the weights are calculated by regressing the forecasts against the actual values i.e. the realized variance. The coefficients from the regression serve as weights. The regression includes an intercept which adjust any bias and therefore it is not necessary that the individual forecasts are unbiased.

The last two methods involve the mean squared error. Stock and Watson (2001) propose MSE weighting in their paper where they compare models for forecasting macroeconomic time series. The MSEs are computed and forecast weights are calculated as

$$\omega_i = \frac{1/MSE_i^k}{\sum_{j=1}^{N} 1/MSE_j^k}$$

where k is set to one. A similar and also last method is the MSE Ranks method introduced by Aiolfi and Timmermann (2006). In this method, the MSEs are ranked and the forecast models are weighted inversely to their rank.

6. EMPIRICAL RESULTS AND DISCUSSION

6.1 INDIVIDUAL FORECASTS

This section presents the out-of-sample predictive power at different horizons of the individual models described in section 2. Table 7 shows the MSE of all models when forecasting volatility one day ahead, one month, six months, one year and two years ahead. The lowest MSEs are indicated by bold figures. A lot of conclusions can be drawn from this table. When forecasting tomorrow's volatility of, the ARMA(1,1) model provides the best forecasts for both indices and for forecasting volatility in one month, it is best to use the EWMA model or the SMA with a twelve month lookback for the S&P500 and the IPC respectively. Focussing on the S&P500, one can see that for all the longer horizons EGARCH(1,1) with normally distributed errors has the lowest MSE compared to the other models. The MSEs of ARCH(1) and GARCH(1,1) with Student's *t*-distributed errors were calculated as well, but horizons longer than one month created very high MSEs so they are omitted because there is no doubt about whether to use these models. GARCH(1,1) with generalized error distributed errors shows some very high MSEs as well and again especially for the longer horizons. Even though the in-sample data showed high values for skewness, these results can be quite confusing. Wilhelmsson (2006) provides a possible solution for this by examining that only a few outliers are the cause of the observed skewness and those outliers have a huge positive impact on the log-likelihood of models that allow for skewness. This is very likely since the S&P500 shows larger outliers than the IPC. It coincides with Christoffersen (2012) too who states that when using a large enough sample (i.e. 1,000 observations) the distribution of the errors does not matter anymore and the normal distribution can be used in order to get the most accurate forecasts. It corresponds to Brownlees et al. (2012) as well who neither find any evidence for using other distributions other than the normal distribution.

The MSEs of ARCH(1) and GARCH(1,1) with Student's *t*-distributed errors of the IPC are included because they fall within 'acceptable' ranges, but they are still very high at the longer horizons relative to the other MSEs. One can also see that at the shorter horizons, HIS models provide lower MSEs than any of the ARCH models. A difference between the S&P500 and the IPC is that GJR-GARCH under the normal distribution shows the lowest MSE at the one and two year horizon instead of EGARCH. Another difference can be found when looking at the overall values. The MSEs of the IPC are lower than those of the S&P500 at all horizons. Probably because the volatility peaks take less time to recover. For all models, the MSE increases as the forecast horizons gets longer which is different than Figlewski (2004) stated in his paper. This could be due to him using another period of time (1947-1995), because these results come from using daily data instead of using monthly data or because he uses another measure of RV as benchmark which is probably much more noisy as the RV used in this paper. It does coincides with Brownlees et al. (2012).

Table 7: MSEs of all models at different horizons.
Panel A: S&P 500

	One-day	One month	Six months	One year	Two years
	ahead	ahead	ahead	ahead	ahead
SMA 6 months	6.607	7.727	9.851	10.738	12.514
SMA 12 months	7.260	7.901	9.269	10.135	11.070
EWMA	4.343	**6.911**	10.903	12.061	14.278
ARMA (1,1)	**3.107**	8.897	10.388	11.562	12.462
ARCH(1,1) - normal distribution	6.430	8.458	9.356	9.956	10.869
ARCH(1,1) - generalized error distribution	6.221	8.579	9.353	9.920	10.830
GARCH(1,1) - normal distribution	3.928	7.428	10.212	10.547	11.579
GARCH(1,1) - generalized error distribution	4.023	8.235	14.958	18.908	29.511
EGARCH(1,1) - normal distribution	6.074	8.086	**8.864**	**9.197**	**10.161**
EGARCH(1,1) - student's t distribution	6.604	8.107	8.973	9.307	10.230
EGARCH(1,1) - generalized error distribution	5.987	8.111	8.963	9.288	10.216
GJR-GARCH(1,1) - normal distribution	3.546	7.959	9.129	9.448	10.422
GJR-GARCH(1,1) - student's t distribution	3.692	8.423	9.549	9.970	10.975
GJR-GARCH(1,1) - generalized error distribution	3.501	8.202	9.412	9.769	10.702

Panel B: IPC

	One-day	One month	Six months	One year	Two years
	ahea	ahead	ahead	ahead	ahead
SMA 6 months	3.460	3.748	4.140	4.507	5.194
SMA 12 months	3.559	**3.729**	**4.063**	4.418	4.987
EWMA	2.897	3.838	4.520	4.893	5.805
ARMA (1,1)	**2.374**	4.175	4.391	4.724	5.374
ARCH(1,1) - normal distribution	3.771	4.715	5.040	5.330	5.853
ARCH(1,1) – student’s t distribution	4.454	7.003	7.558	7.950	8.787
ARCH(1,1) - generalized error distribution	3.703	4.658	4.972	5.258	5.769
GARCH(1,1) - normal distribution	3.307	4.171	4.891	5.747	6.974
GARCH(1,1) – student’s t distribution	3.361	4.311	5.355	7.082	11.53
GARCH(1,1) - generalized error distribution	3.308	4.186	4.796	5.513	6.474
EGARCH(1,1) - normal distribution	3.051	3.852	4.305	4.444	4.823
EGARCH(1,1) - student's t distribution	3.053	3.863	4.333	4.483	4.798
EGARCH(1,1) - generalized error distribution	3.053	3.867	4.335	4.474	4.835
GJR-GARCH(1,1) - normal distribution	2.926	3.923	4.253	4.383	4.735
GJR-GARCH(1,1) - student's t distribution	2.920	3.914	4.235	**4.367**	**4.698**
GJR-GARCH(1,1) - generalized error distribution	2.923	3.975	4.294	4.427	4.762

MSEs are calculated as $\frac{1}{n}\sum_{t=1}^{n}(\hat{\sigma}_t - \sigma_t)^2$ with 5-minute intraday squared returns as proxy for actual volatility. n is 3,352, 3,332, 3,227,3,101 and 2,849 for the one day, one month, six months, one year and two years horizon respectively.

Besides comparing the MSEs in table 7, it is important to compare the models in a statistical way using the Diebold Mariano (1995) test discussed in the previous section. Because comparing each model as a pair is not the most appropriate method when evaluating multiple models, The Model Confidence Set by Hansen, Lunde and Nason (2011) is an easier way to determine the 'best' forecasts at each horizon. Still the Diebold Mariano (1995) test statistics for the S&P500 are reported in appendix C to show conclusions are the same. The test statistics are calculated in two ways but the results coincide, so only the regression-based statistics will be reported in the appendix.

According to table 7 panel A, the ARMA(1,1) model outperforms all other models based on their MSEs. However when looking at the Diebold Mariano test statistics, there is no significant difference between the loss function of the ARMA(1,1) model and any of the other models except for both SMA. For the one day ahead forecast it becomes clear that almost all models are significantly better than the simple moving averages. When comparing all other models, there is no overall conclusion that can be made. All models perform quite the same and there are not many significant p-values. What can be noted is that GARCH(1,1), normally distributed, outperforms the ARCH(1) models and that GJR-GARCH under a generalized error distribution outperforms many models as well. At the one month horizon, EWMA shows the lowest MSE. However, when looking at the test statistics, EWMA is only significant better than SMA and EWMA at the 5% level. Another outcome is that the ARCH(1) model under a generalized error distribution is significantly worse than seven other models. At the one day ahead forecast, both simple moving averages were outperformed by almost all models. At a longer horizon, i.e. one month, this is no longer the case. Maybe because variance tend to move towards its average value in the longer term.

Where it is hard to see one model being better than others at the shorter horizons, at the longer horizons there is a clear pattern. First of all, using the GARCH(1,1) model under a generalized error distribution is probably not a good idea since it performs statistically worse than all other models at all horizons longer than one month. The same holds for ARMA(1,1), which is outperformed by almost all other models as well. For six months, one year and two year horizons it is best to forecast volatility with a normally distributed EGARCH(1,1) model as evidenced by table 6. This confirmed by the Diebold Mariano test as well. At the one and two year horizon this model performs significantly better than all other models. At the six months horizon it outperforms all models except for the simple moving averages.

An overall striking result is that moving averages i.e. the historical volatility models, perform not as badly as stated in the existing literature. Perhaps this is not so surprising since there is a lot of improvement in creating realized data which is used as an input in the moving averages. The simple moving averages only fail at forecasting tomorrow's volatility or the volatility in two years. For all other horizons it has equal predictive ability as all other models except for EGARCH and GJR-GARCH.

The main outcome is that EGARCH performs significantly best at the longer horizons, especially under the normal distribution. At the shorter horizons however, EGARCH does not perform outstanding at all. At the one month horizon it is outperformed by both ARCH models and compared to all other models its predictive ability is not statistically better or worse. When looking at the goodness of fit test done in section 5, EGARCH also showed the lowest AIC of all. So a good fit of the in-sample data might suggest a higher predictive ability as well.

Table 8 panel A reports the MSE and MCS p-values for each of the individual forecasts of the S&P500 volatility. The models included in $\mathcal{M}^*$ are identified by an asterisk. At the one month horizon only three models are excluded from the set so it is not possible to determine a few models which perform best. At the one day horizon, half of the models is included in the set: the GARCH models, GJR-GARCH models, EWMA and ARMA. At almost every horizon the ARCH model is excluded from the set. At the long horizon, only the EGARCH model under the normal distribution is included in the set. Even for different block lengths and/or bootstrap replications these results stay the same. This is some more evidence that at the longer horizons it is best to use the EGARCH model under the normal distribution. EGARCH under the normal distribution belongs to the set at every other horizon as well, except for the one day ahead.

Panel B reports the results of the IPC. Again, the ARCH model under all three different distributions is excluded from the set. At the two year horizon there is a clear-cut result that the GJR-GARCH performs best since $\mathcal{M}^*$ only consists of GJR-GARCH models. At all other horizons except for the one month horizon, the set consists of all three GJR-GARCH models as well. EGARCHN performs well for the IPC as well. It is included in all other sets except the set of best performing models at the two years horizon. The fact that GJR-GARCH performs slightly better could be due to the fact that GJR-GARCH can change volatility quite fast when the return changes sign. This is some feature we saw in figure 2.

Table 8: MCS p-values for individual volatility forecasts at different forecast horizons.
Panel A: S&P 500

One day ahead		One month ahead		Six months ahead		One year ahead		Two years ahead	
Model	P_{MCS}	**Model**	P_{MCS}	**Model**	P_{MCS}	**Model**	P_{MCS}	**Model**	P_{MCS}
SMA12	0.000	ARMA	0.017	GARCHGED	0.046	GARCHGED	0.029	GARCHGED	0.034
SMA6	0.000	ARCHGED	0.017	ARMA	0.046	ARMA	0.029	EWMA	0.034
EGARCHT	0.000	GJRGARCHT	0.017	GARCHN	0.046	GARCHN	0.029	ARMA	0.034
ARCHN	0.000	ARCHN	0.11*	EWMA	0.046	EWMA	0.029	SMA6	0.034
ARCHGED	0.000	GJRGARCHGED	0.11*	GJRGARCHT	0.046	SMA6	0.029	GJRGARCHT	0.034
EGARCHN	0.000	EGARCHGED	0.11*	GJRGARCHGED	0.046	GJRGARCHT	0.029	GARCHN	0.034
EGARCHGED	0.048	EGARCHT	0.51*	SMA6	0.05*	ARCHN	0.029	GJRGARCHGED	0.034
EWMA	0.10*	EGARCHN	0.51*	ARCHN	0.05*	GJRGARCHGED	0.029	SMA12	0.034
GARCHGED	0.10*	GARCHGED	0.51*	ARCHGED	0.05*	ARCHGED	0.029	ARCHN	0.034
GJRGARCHT	0.10*	GJRGARCHN	0.59*	GJRGARCHN	0.05*	SMA12	0.029	ARCHGED	0.034
GARCHN	0.30*	SMA12	0.59*	SMA12	0.05*	GJRGARCHN	0.029	GJRGARCHN	0.034
GJRGARCHN	0.73*	SMA6	0.59*	EGARCHT	0.05*	EGARCHT	0.029	EGARCHT	0.034
GJRGARCHGED	0.73*	GARCHN	0.59*	EGARCHGED	0.05*	EGARCHGED	0.029	EGARCHGED	0.034
ARMA	1.00*	EWMA	1.00*	EGARCHN	1.00*	EGARCHN	1.00*	EGARCHN	1.00*

Panel B: IPC

One day ahead		One month ahead		Six months ahead		One year ahead		Two years ahead	
Model	P_{MCS}	**Model**	P_{MCS}	**Model**	P_{MCS}	**Model**	P_{MCS}	**Model**	P_{MCS}
ARCHT	0.000	ARCHT	0.000	ARCHT	0.009	ARCHT	0.011	ARCHT	0.004
ARCHN	0.000	ARCHN	0.000	GARCHT	0.009	GARCHT	0.011	GARCHT	0.004
ARCHGED	0.000	ARCHGED	0.000	ARCHN	0.009	ARCHN	0.011	GARCHN	0.004
SMA12	0.000	ARMA	0.003	ARCHGED	0.009	GARCHN	0.011	GARCHGED	0.004
GARCHT	0.000	GARCHT	0.003	GARCHN	0.009	ARCHGED	0.011	ARCHN	0.004
GARCHGED	0.001	GARCHGED	0.003	GARCHGED	0.009	GARCHGED	0.011	ARCHGED	0.004
SMA6	0.001	GARCHN	0.003	EWMA	0.009	ARMA	0.011	ARMA	0.004
GARCHN	0.08*	GJRGARCHGED	0.003	EGARCHGED	0.009	EWMA	0.011	EWMA	0.004
EGARCHT	0.39*	GJRGARCHN	0.003	EGARCHT	0.021	EGARCHT	0.011	SMA6	0.004
EGARCHGED	0.39*	GJRGARCHT	0.003	ARMA	0.24*	EGARCHGED	0.024	EGARCHGED	0.004
EGARCHN	0.39*	EGARCHGED	0.003	EGARCHN	0.24*	SMA6	0.15*	SMA12	0.004
GJRGARCHGED	0.39*	EGARCHT	0.11*	GJRGARCHGED	0.84*	EGARCHN	0.15*	EGARCHN	0.004
GJRGARCHN	0.39*	EGARCHN	0.62*	GJRGARCHN	0.89*	GJRGARCHGED	0.81*	EGARCHT	0.026
GJRGARCHT	0.39*	EWMA	0.85*	GJRGARCHT	0.89*	SMA12	0.87*	GJRGARCHGED	0.72*
EWMA	0.39*	SMA6	0.85*	SMA6	0.89*	GJRGARCHN	0.87*	GJRGARCHN	0.72*
ARMA	1.00*	SMA12	1.00*	SMA12	1.00*	GJRGARCHT	1.00*	GJRGARCHT	1.00*

The model confidence set algorithm uses a block-bootstrap procedure with 1,000 bootstrap replications and at the one day ahead horizon a block length of 2 is used, at the one month horizon, six months horizon, one year horizon and two year horizon, a block length of 21, 126, 252 and 504 is used respectively. The models included in $\mathcal{M}^*$ are identified by an asterisk. The MCS p-value for model $e_{\mathcal{M}_j} \in \mathcal{M}^0$ is defined by $\hat{p}_{e_{\mathcal{M}_j}} = max_{i \leq j} P_{H_{0,\mathcal{M}_i}}$.

6.2 FORECAST COMBINATIONS

Besides identifying the single best forecasting model at each horizon, it might be better to combine forecasts in order to get lower MSEs. Table 9 shows MSEs of different methods to combine forecasts of the S&P500. All these loss functions are compared to the individual best performing model (shown in the last column) with the Diebold Mariano (1995) test to see whether they have indeed superior predictive power or not.

Tabel 9: MSEs of forecast combinations at different horizons for the S&P 500.

Panel A: One day ahead forecasts

	Simple Mean	Trimmed Mean (10%)	Simple Median	Least Squares	MSE Weights	MSE Ranks	Individual
MSE	3.652	3.853	4.032	4.822	3.285	**3.064**	3.107

Panel B: One month ahead forecasts

	Simple Mean	Trimmed Mean (10%)	Simple Median	Least Squares	MSE Weights	MSE Ranks	Individual
MSE	6.754	6.860	7.397	7.605	6.723	**6.604**	6.911

Panel C: Six months ahead forecasts

	Simple Mean	Trimmed Mean (30%)	Simple Median	Least Squares	MSE Weights	MSE Ranks	Individual
MSE	8.460	**8.354**	8.425	11.651	8.407	8.391	8.864

Panel D: One year ahead forecasts

	Simple Mean	Trimmed Mean (85%)	Simple Median	Least Squares	MSE Weights	MSE Ranks	Individual
MSE	9.116	**8.832**	**8.832**	9.197	8.995	8.897	9.197

Panel E: Two years ahead forecasts

	Simple Mean	Trimmed Mean (55%)	Simple Median	Least Squares	MSE Weights	MSE Ranks	Individual
MSE	10.284	9.855	9.890	9.950	10.010	**9.839**	10.161

For each horizon, another level of trimming is used which is reported in brackets. The last column reports the lowest MSE of the individual forecasts for comparison. The lowest MSEs are indicated by bold figures. Different horizons are separated by panel A, B, C, D and E.

What becomes clear is that at every horizon, at least one of the forecast combinations provides a lower MSE than the lowest MSE of the best individual forecast, especially at the one and two year ahead forecasts. MSE ranks provides the lowest MSE in three out of five horizons. Timmermann (2006) concluded that the trimmed mean often improves performance as well. The trimmed mean combination weights are calculated with different levels of trimming. The optimal percentage is given in the table as well. These results suggest that trimming indeed improves performance except for the shorter horizons. This might be due to the fact that there is not much of a difference between the predictive power of the models in forecasting volatility in the short run.

Table 11 provides some more insights on whether the best performing forecast combination, statistically beats the best performing individual forecast. EGARCH is outperformed by the forecast combinations at the longer horizons. At the shorter horizons, the individual best performing model is not beaten by the forecast combinations.

Table 11: Diebold Mariano Test (1995) T-statistics S&P500 forecast combinations versus individual forecasts.

	One day ahead ARMA	One month ahead EWMA	Six months ahead EGARCHN	One year ahead EGARCHN	Two years ahead EGARCHN
MSE Ranks	-0.042	-0.544			-3.914***
Trimmed Mean			-4.600***	-3.731***	

The left column shows the forecast combinations methods with the lowest MSE at each horizon. For the six months and one year horizon the trimmed mean provides the lowest MSE and for the other horizons MSE Ranks performs best. These MSEs are compared with the individual best performing methods. *** indicates significance at the 1% level.

Table 10 shows the results for the IPC. Again MSE Ranks shows the lowest MSEs in four out of five times. Also, at none of these horizons, the individual forecasts provide the lowest MSE no more. Table 12 shows the Diebold Mariano (1995) test statistics and only ARMA is statistically outperformed. All other forecast combinations are not statistically better than the individual forecast.

Tabel 10: MSEs of forecast combinations at different horizons for the IPC.

Panel A: One day ahead forecasts

	Simple Mean	Trimmed Mean (10%)	Simple Median	Least Squares	MSE Weights	MSE Ranks	Individual
MSE	2.523	2.548	2.709	**1.949**	2.475	2.377	2.374

Panel B: One month ahead forecasts

	Simple Mean	Trimmed Mean (30%)	Simple Median	Least Squares	MSE Weights	MSE Ranks	Individual
MSE	3.620	3.584	3.600	3.662	3.585	**3.548**	3.729

Panel C: Six months ahead forecasts

	Simple Mean	Trimmed Mean (75%)	Simple Median	Least Squares	MSE Weights	MSE Ranks	Individual
MSE	4.006	3.910	3.920	4.344	3.959	**3.906**	4.063

Panel D: One year ahead forecasts

	Simple Mean	Trimmed Mean (50%)	Simple Median	Least Squares	MSE Weights	MSE Ranks	Individual
MSE	4.288	4.170	4.207	4.876	4.207	**4.130**	4.367

Panel E: Two years ahead forecasts

	Simple Mean	Trimmed Mean (50%)	Simple Median	Least Squares	MSE Weights	MSE Ranks	Individual
MSE	4.829	4.661	4.777	5.869	4.673	**4.558**	4.698

For each horizon, another level of trimming is used which is reported in brackets. The last column reports the lowest MSE of the individual forecasts for comparison. The lowest MSEs are indicated by bold figures. Different horizons are separated by panel A, B, C, D and E.

Table 12: Diebold Mariano Test (1995) T-statistics IPC forecast combinations versus individual forecasts.

	One day ahead ARMA	One month ahead SMA12	Six months ahead SMA12	One year ahead GJR-GARCHT	Two years ahead GJR-GARCHT
MSE Ranks		-0.105	-0.086	-0.174	-0.460
Least Squares	-2.815***				

The left column shows the forecast combinations methods with the lowest MSE at each horizon. For the six months and one year horizon the trimmed mean provides the lowest MSE and for the other horizons MSE Ranks performs best. These MSEs are compared with the individual best performing methods. *** indicates significance at the 1% level.

7. SUMMARY AND CONCLUSIONS

This study focusses on the forecasting performance at multiple horizons. First the data was analysed too see if the theory about stylized facts hold for the data used in this paper as well. In-sample tests showed that both series are stationary at the 1% level, volatility clustering exists and volatility is definitely time-varying. Engle and Ng's (1993) sign and bias test showed there is an asymmetry or leverage effect as well. The negative size bias is present in the S&P500 time series and both negative and positive size bias is present in the IPC returns. Besides this, both time series coincide with theory about the presence of conditional heteroscedasticity as well.

A large part of the existing literature shows that financial time series data exhibit skewness and excess kurtosis as well as the Jarque-Bera test suggested in this study. This has been taken into account by using a Student's *t*- and a Generalized Error distribution in addition to the normal distribution. Unfortunately this does not improve the accuracy of the forecasts. At every horizon and for every model it is best to use the normal distribution. Especially forecasting volatility of the S&P500 with models under the Student's *t*- or Generalized Error distribution provides very high values of MSE.

The applied models can be classified into historical volatility models: SMA, EWMA and ARMA(1,1) which uses conditional variance as inputs and regression-based models containing ARCH(1), GARCH(1,1), EGARCH(1,1) and GJR-GARCH. To evaluate those individual models, the loss function MSE is used and intraday 5-minute squared returns as proxy for actual volatility. The preferred models have been selected by applying the Diebold Mariano (1995) test and the Model Confidence Set. One general conclusion is that the MSE increases as the forecast horizon lengthens.

At the one day horizon it is hard to draw conclusions. A lot of models seem to produce statistically the same loss functions. The MCS algorithm includes more than half of the evaluated models in the best set of both indices. For the S&P500 the result at the longer horizons is very clear-cut. The best set only includes EGARCH under the normal distribution. No other model performs equally. For the IPC it is the GJR-GARCH model under all distributions that is included solely in the best set at the two year horizon. The results coincide with existing literature that asymmetric models seem to perform best and that historical models definitely provide accurate forecasts as well.

Besides the individual forecasts, different forecast combinations were evaluated as well. Between the different methods, MSE Ranks provided the lowest MSEs in most cases. The Diebold Mariano test showed that statistically there is not much of a difference between choosing the best individual forecast in isolation or to combine forecasts. Except for the S&P500. The test statistics show that the MSE Ranks and the Trimmed Mean produce statistically lower MSEs than EGARCH under the normal distribution which was initially the best performing individual model.

Research shows that is it fruitful to incorporate the stylized facts. Especially models that take the leverage effect into account seem to perform well. Based on this results, GJR-GARCH works well for return series that show very short volatile periods like the IPC. EGARCH works better for the S&P500 index which exhibits slowly decaying volatility. The profits from taking stylized facts into account come forward especially at the longer horizons. At the shorter horizons it is hard to draw conclusions except that simple historical models suffice and models that perform well in the long run are not a guarantee that they also perform well in the short run. Using other distributions than the normal distribution is not recommended but for further research I would suggest to use other loss functions as well since this might highlight the benefits of using different distribution than the normal distribution.

REFERENCES

Aiolfi, M., & Timmermann, A. (2006). Persistence in Forecasting Performance and Conditional Combination Strategies. *Journal of Econometrics, 135*(1), 31-53.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *2nd Internation Symposium on Information Theory*, 267-281.

Akgiray, V. (1989). Conditional Heteroskedasticity in Time Series of Stock Returns: Evidence and Forecasts. *Journal of Business, 62*, 55-80.

Alford, A., & Boatsman, J. (1995). Predicting Long-Term Stock Return Volatility: Implications for Accounting and Valuation of Equity Derivatives. *The Accounting Review, 70*(4), 559-618.

Andersen, T. G., & Bollerslev, T. (1998). Deutsche Mark-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies. *The Journal of Finance, 53*(1), 219-265.

Brailsford, T., & Faff, R. (1996). An Evaluation of Volatility Forecasting Techniques. *Journal of Banking and Finance, 20*(3), 419-438.

Brownlees, C., Engle, R., & Kelly, B. (2012). A Practical Guide to Volatility Forecasting through Calm and Storm. *The Journal of Risk, 14*(2), 3-22.

Cao, C., & Tsay, R. (1992). Nonlinear Time-Series Analysis of Stock Volatilities. *Journal of Applied Econometrics, 12*(1), 165-185.

Christie, A. A. (1982). The Stochastic Behavior of Common Stock Variances. *Journal of Financial Economics, 10*, 407-432.

Christoffersen, P. F. (2012). *Elements of Financial Risk Management* (Vol. 2). Oxford: Elsevier.

Dickey, W. A., & Fuller, D. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association, 74*, 427-431.

Engle, R. F., & Ng, V. K. (1993). Measuring and Testing the Impact of News on Volatility. *The Journal of Finance, 48*(5), 1749-1778.

Figlewski, S. (1997). Forecasting Volatility. *Financial Markets, Institutions & Instruments, 6*(1), 1-88.

Figlewski, S., & Green, T. C. (1999). Market Risk and Model Risk for a Financial Institution Writing Options. *The Journal of Finance, 54*(4), 1465-1499.

Glosten, L., R.Jagannathan, & Runkle, D. (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Returns on Stocks. *The journal of finance, 48*(5), 1779-1801.

Granger, C. W., & Poon, S. H. (2003). Forecasting Volatility in Financial Markets: A Review. *Journal of Economic Literature, 41*(2), 478-539.

Hansen, P. R., & Lunde, A. (2005). A Forecast Comparison of Volatility Models: Does anything Beat a GARCH(1,1)? *Journal of Applied Econometrics, 20*, 873-889.

Hansen, P., Lunde, A., & Nason, J. (2011). The Model Confidence Set. *Econometrica, 79*(2), 453-497.

Knight, J., & Satchell, S. (2007). *Forecasting Volatility in the Financial Markets* (3rd Edition ed.). London: Elsevier.

Liu, L. Y., Patton, A. J., & Sheppard, K. (2015). Does Anything Beat 5-minute RV? A Comparison of Realized Measures across Multiple Asset Classes. *Journal of Econometrics, 187*(1), 293-311.

Makridakis, S., & Winkler, R. (1983). Averages of Forecasts: Some Empirical Results. *Management Science, 29*(9), 987-996.

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica, 59*(2), 347-370.

Pagan, A., & Schwert, G. (1990). Alternative Models for Conditional Models for Conditional Stock Volatility. *Journal of Econometrics, 45*(1-2), 267-290.

Poon, S. H. (2005). *A Practical Guide to Forecasting Financial Market Volatility.* Chichester: John Wiley & Sons, Ltd.

Quaedvlieg, R. (2017). Multi-Horizon Forecast Comparison. *Erasmus School of Economics, Erasmus University Rotterdam*, 1-35.

Stock, J., & Watson, M. (2001). A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series. *Cambridge University Press*, 1-44.

Taylor, J. (2004). Volatility Forecasting with Smooth Transition Exponential Smoothing. *International Journal of Forecasting, 20*, 273-286.

Timmermann, A. (2006). Forecast Combinations. *Handbook of economic forecasting, 1*, 135-196.

Tsay, R. (2002). *Analysis of Financial Time Series: Financial Econometrics.* Chichester: John Wiley & Sons Ltd.

Wilhelmsson, A. (2006). GARCH Forecasting Performance under Different Distribution Assumptions. *Journal of Forecasting, 25*(8), 561-578.

APPENDIX A: REALIZED VOLATILITY

Figure A1: Realized volatility S&P500 from 1/26/2004 to 6/5/2008

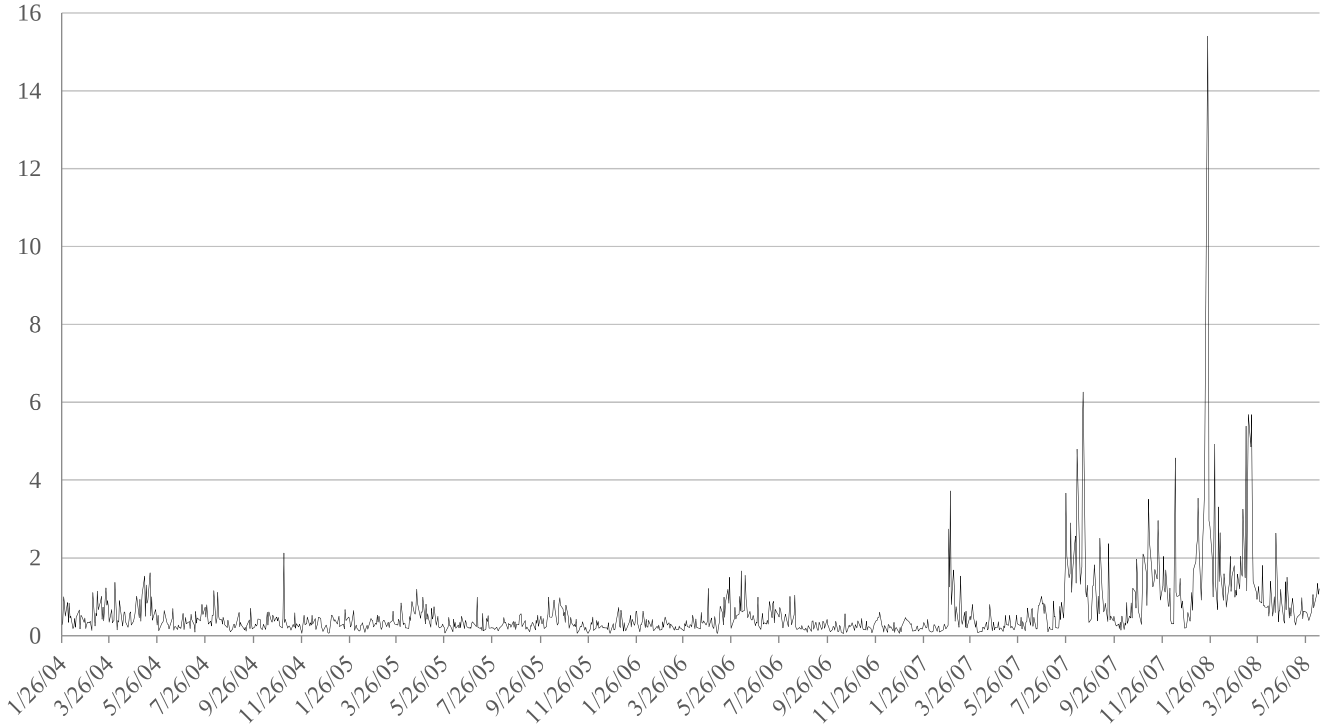


Figure A2: Realized volatility IPC from 1/26/2004 to 6/5/2008

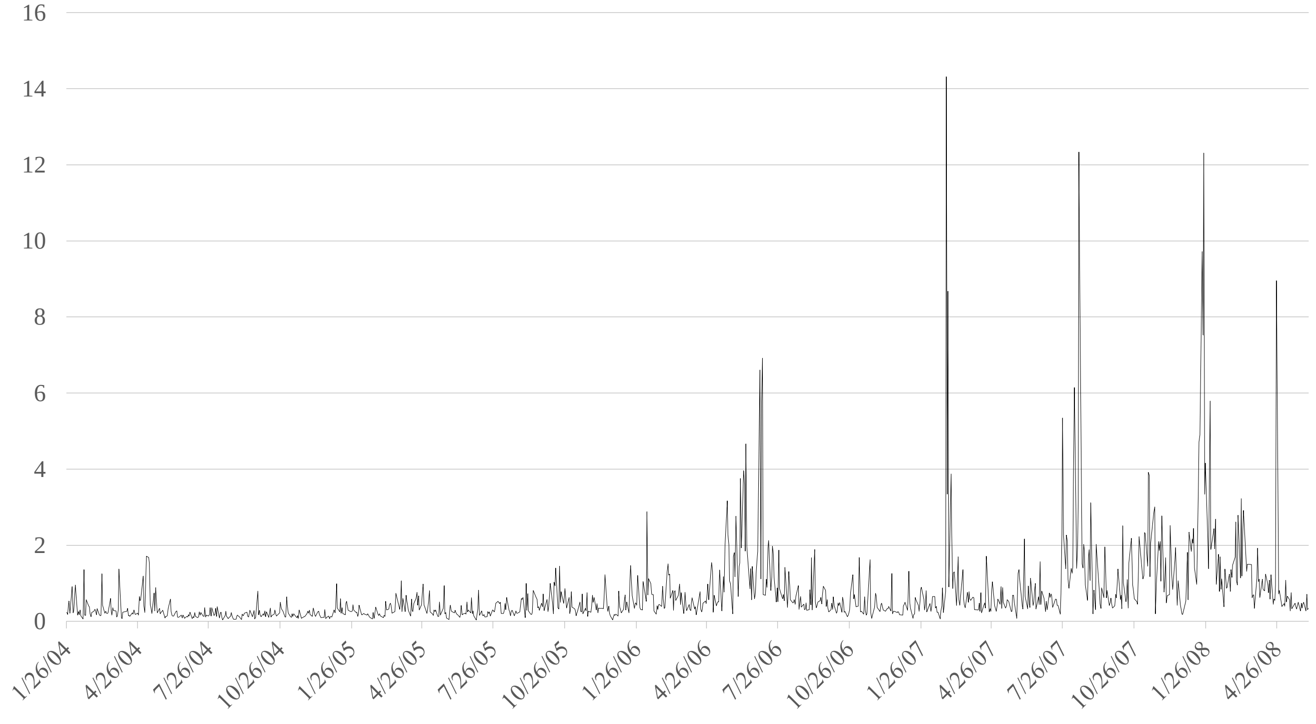


Figure A3: Realized Volatility S&P500 6/9/2008 to 10/31/2012

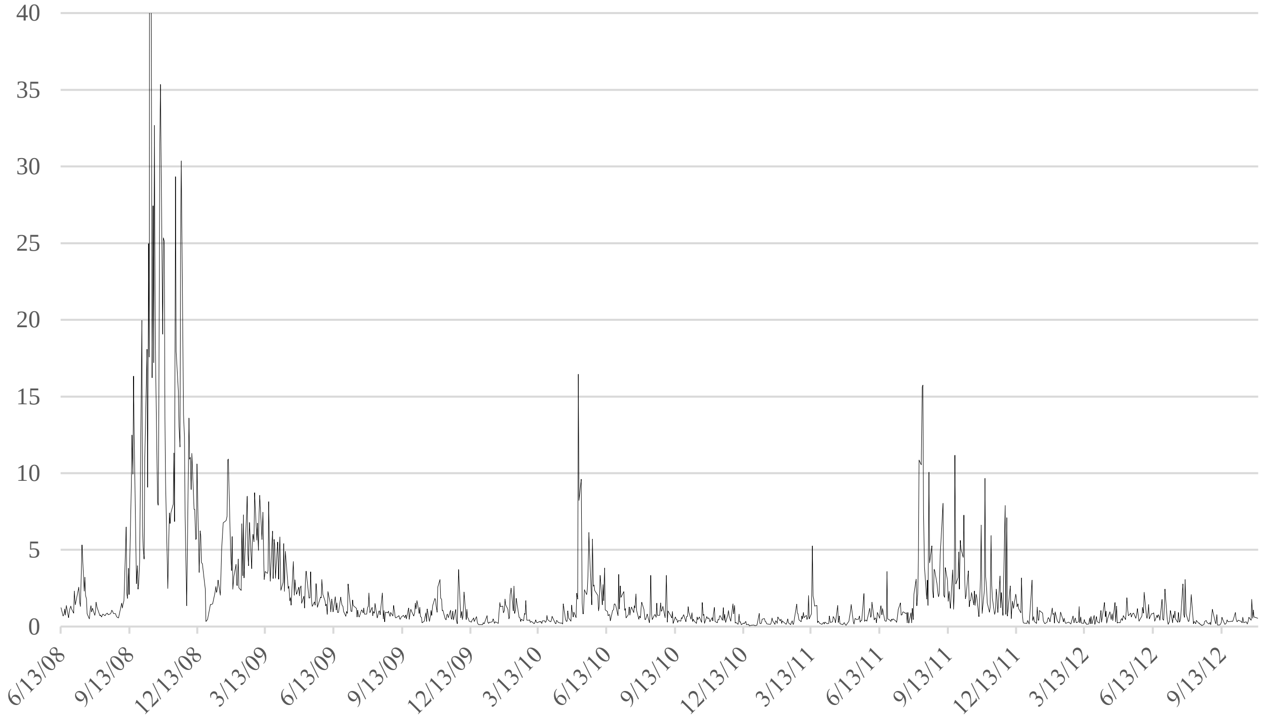


Figure A4: Realized Volatility IPC 6/9/2008 to 10/31/2012

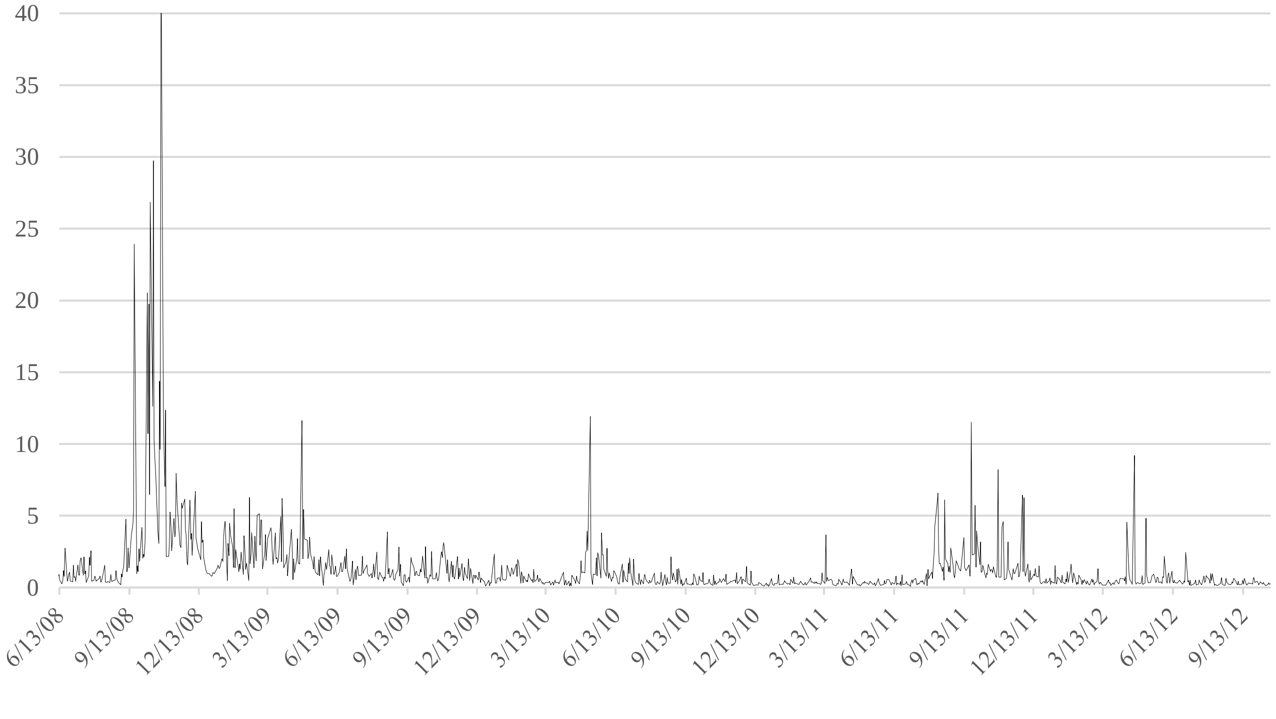


Figure A5: Realized Volatility S&P500 11/1/2012 to 5/31/2017

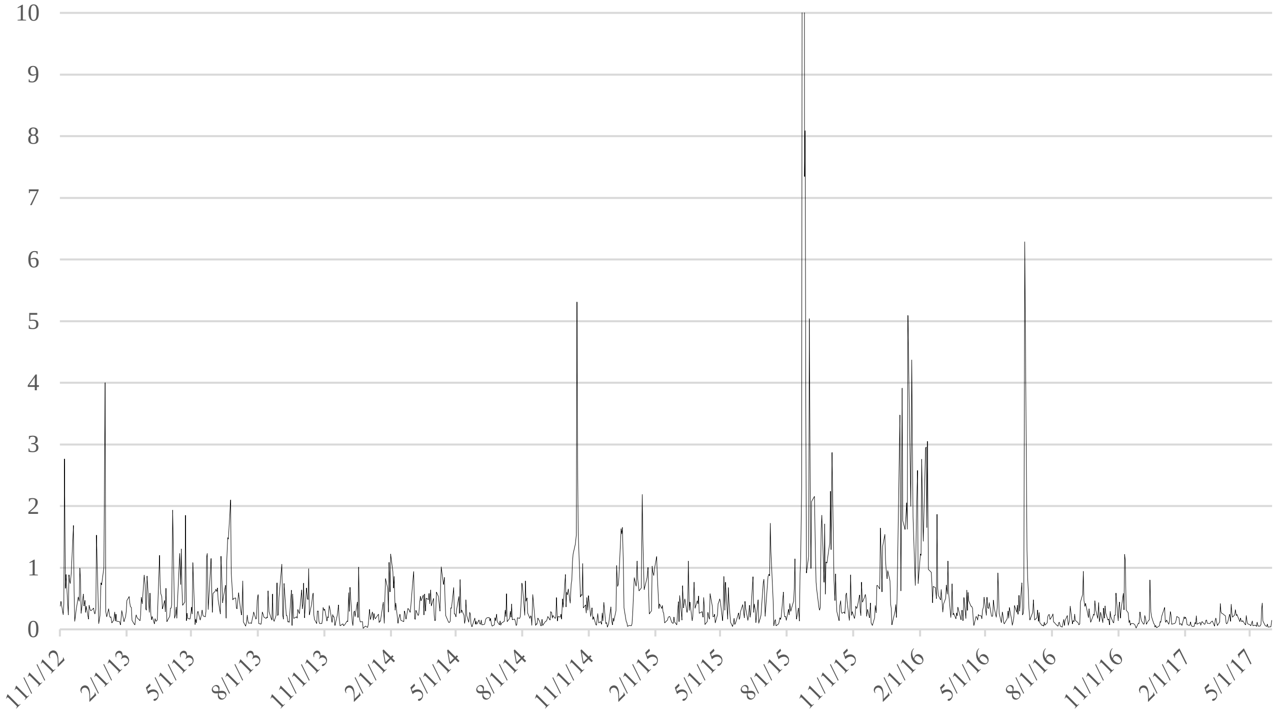
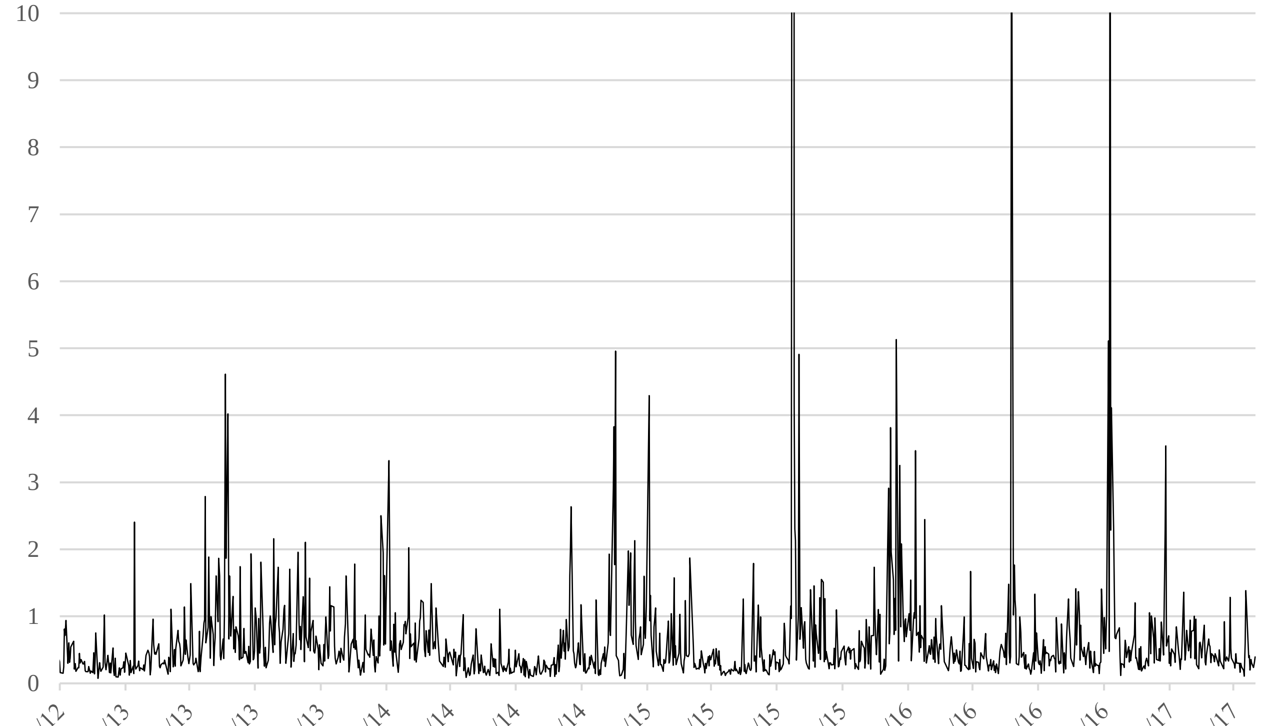


Figure A6: Realized Volatility IPC 11/1/2012 to 5/31/2017



APPENDIX B: IN-SAMPLE PARAMETER ESTIMATES

S&P500

Table B1: In-sample parameter estimates of all models forecasting volatility of the S&P500 returns.

ARMA (1,1)	Estimate
α_0	1.502
α_j	0.926
β_j	-0.613

GARCH(1,1) Normal	Estimate	GARCH(1,1) *t*-dist	Estimate	GARCH(1,1) GED	Estimate
ω	0.027	ω	0.027	ω	0.026
α_j	0.080	α_j	0.079	α_j	0.079
β_j	0.906	β_j	0.907	β_j	0.907

ARCH(1) Normal	Estimate	ARCH(1) *t*-dist	Estimate	ARCH(1) GED	Estimate
ω	1.573	ω	1.648	ω	1.605
α_j	0.157	α_j	0.120	α_j	0.132

EGARCH(1,1) Normal	Estimate	EGARCH(1,1) *t*-dist	Estimate	EGARCH(1,1) GED	Estimate
α_0	-0.042	α_0	-0.041	α_0	-0.042
α_1	0.060	α_1	0.058	α_1	0.059
θ	-0.129	θ	-0.134	θ	-0.131
γ	0.980	γ	0.980	γ	0.980

GJRGARCH(1,1) Normal	Estimate	GJRGARCH(1,1) *t*-dist	Estimate	GJRGARCH(1,1) GED	Estimate
ω	0.027	ω	0.026	ω	0.026
α_j	0.019	α_j	0.027	α_j	0.023
δ_j	0.170	δ_j	0.170	δ_j	0.169
β_j	0.922	β_j	0.930	β_j	0.926

IPC

Table B2: In-sample parameter estimates of all models forecasting volatility of the S&P500 returns.

ARMA (1,1)	Estimate
α_0	0.896
α_j	0.997
β_j	-0.957

GARCH(1,1) Normal	Estimate	GARCH(1,1) *t*-dist	Estimate	GARCH(1,1) GED	Estimate
ω	0.013	ω	0.017	ω	0.014
α_j	0.045	α_j	0.054	α_j	0.049
β_j	0.947	β_j	0.937	β_j	0.943

ARCH(1) Normal	Estimate	ARCH(1) *t*-dist	Estimate	ARCH(1) GED	Estimate
ω	1.754	ω	1.856	ω	1.746
α_j	0.285	α_j	0.300	α_j	0.277

EGARCH(1,1) Normal	Estimate	EGARCH(1,1) *t*-dist	Estimate	EGARCH(1,1) GED	Estimate
α_0	-0.043	α_0	-0.068	α_0	-0.058
α_1	0.0640	α_1	0.100	α_1	0.085
θ	-0.080	θ	-0.080	θ	-0.079
γ	0.983	γ	0.980	γ	0.981

GJRGARCH(1,1) Normal	Estimate	GJRGARCH(1,1) *t*-dist	Estimate	GJRGARCH(1,1) GED	Estimate
ω	0.021	ω	0.030	ω	0.024
α_j	0.010	α_j	0.008	α_j	0.000
δ_j	0.073	δ_j	0.085	δ_j	0.077
β_j	0.959	β_j	0.932	β_j	0.946

APPENDIX C: DIEBOLD MARIANO TEST STATISTICS

Table C1: Diebold Mariano (1995) test statistics of the S&P500 comparing all models at the one day horizon.

	SMA 6	SMA 12	EWMA	ARMA	ARCHN	ARCHG	GARCHN	GARCHG	EGARCHN	EGARCHT	EGARCHG	GJRGARCHN	GJRGARCHT	GJRGARCHG
SMA 6	-	-2.145*	2.43**	1.518	0.412	1.071	2.234**	2.071*	1.723	0.007	2.058*	2.196*	2.024*	2.177*
SMA 12		-	2.655*	1.714	1.625	2.172*	2.479*	2.334*	4.132***	1.511	4.260***	2.413**	2.256**	2.393**
EWMA			-	0.801	-2.129*	-2.038*	1.207	0.766	-1.792	-2.312**	-1.757	1.425	1.063	1.431
ARMA(1,1)				-	-1.457	-1.381	-0.634	-0.917	-1.317	-1.566	-1.294	-0.4145	-0.572	-0.387
ARCH(N					-	1.455	2.125*	1.977*	0.811	-0.319	1.025	2.091*	1.928	2.073*
ARCHGED						-	2.004*	1.846	0.361	-0.745	0.591	1.988*	1.821	1.973*
GARCHN							-	-1.041	-1.808	-2.249**	-1.778	1.344	0.728	1.346
GARCHGED								-	-1.673	-2.101*	-1.639	1.711	1.085	1.666
EGARCHN									-	-1.556	2.675**	1.840	1.684	1.831
EGARCHT										-	1.829	2.231*	2.064*	2.215*
EGARCHG											-	1.815	1.656	1.806
GJRGARCHN												-	-1.844	0.709
GJRGARCHT													-	2.201*
GJRGARCHG														-

A regression is run with the difference between column model and row model as dependent variable and a constant as independent variable. T-statistics are reported. Negative values of the Diebold–Mariano test show that the squared errors of the model listed first (column models) are lower than those of the model listed last (row models). ARCH and GARCH under the t-distribution are omitted due to very large MSEs. The significance levels are indicated by *, ** and ***, and correspond to a significance level of 10%, 5% and 1% respectively, using a two-tailed test. Each model produces 3,352 forecasts.

Table C2: Diebold Mariano (1995) test statistics of the S&P500 comparing all models at the one month horizon.

	SMA 6	SMA 12	EWMA	ARMA	ARCHN	ARCHG	GARCHN	GARCHG	EGARCHN	EGARCHT	EGARCHG	GJRGARCHN	GJRGARCHT	GJRGARCHG
SMA 6	-	-0.626	1.169	-2.9**	-1.851	-2.072*	0.318	-0.407	-0.829	-0.887	-0.889	-0.699	-1.931	-1.371
SMA 12		-	1.258	-2.5**	-1.840	-2.122*	0.461	-0.252	-0.497	-0.557	-0.563	-0.198	-1.595	-0.949
EWMA			-	-2.7**	-1.704	-1.811	-1.542	-1.902	-1.239	-1.268	-1.267	-1.256	-1.777	-1.538
ARMA(1,1)				-	1.043	0.727	1.504	0.512	1.539	1.512	1.496	2.184*	1.063	1.582
ARCH(N					-	-3.63***	0,911	0,158	1.967*	1.849	1.827	2.427**	0.151	1.178
ARCHGED						-	1.010	0.244	2.860**	2.709**	2.689**	2.967**	0.681	1.722
GARCHN							-	-2.125*	-0.567	-0.586	-0.588	-0.495	-0.917	-0.720
GARCHGED								-	0.104	0.090	0.087	0.202	-0.137	0.024
EGARCHN									-	-1.655	-3.37***	0.764	-1.736	-0.636
EGARCHT										-	-0.434	0.938	-1.696	-0.548
EGARCHG											-	0.931	-1.641	-0.514
GJRGARCHN												-	-3.769***	-2.451**
GJRGARCHT													-	1.600
GJRGARCHG														-

A regression is run with the difference between column model and row model as dependent variable and a constant as independent variable. T-statistics are reported. Negative values of the Diebold–Mariano test show that the squared errors of the model listed first (column models) are lower than those of the model listed last (row models). ARCH and GARCH under the t-distribution are omitted due to very large MSEs. The significance levels are indicated by *, ** and ***, and correspond to a significance level of 10%, 5% and 1% respectively, using a two-tailed test. Each model produces 3,332 forecasts.

Table C3: Diebold Mariano (1995) test statistics of the S&P500 comparing all models at the six months horizon.

	SMA 6	SMA 12	EWMA	ARMA	ARCHN	ARCHG	GARCHN	GARCHG	EGARCHN	EGARCHT	EGARCHG	GJRGARCHN	GJRGARCHT	GJRGARCHG
SMA 6	-	1.359	-1.111	-0.792	0.858	0.852	-0.551	-2.665**	1.584	1.393	1.408	1.122	0.459	0.684
SMA 12		-	-1.640	-2.23*	-0.283	-0.269	-1.655	-2.868**	1.057	0.754	0.780	0.342	-0.654	-0.346
EWMA			-	0.504	1.510	1.511	1.107	-3.575**	1.995*	1.883	1.893	1.718	1.302	1.447
ARMA(1,1)				-	2.372**	2.387**	0.285	-2.345**	2.880**	2.631**	2.650**	2.300**	1.512	1.777
ARCH(N					-	0.294	-1.675	-2.812**	2.480**	1.829	1.882	1.055	-0.810	-0.248
ARCHGED						-	-1.677	-2.813**	2.445**	1.801	1.853	1.033	-0.822	-0.262
GARCHN							-	-3.170**	2.544**	2.321**	2.341**	2.010**	1.219	1.490
GARCHGED								-	3.059**	3.003**	3.008**	2.922**	2.713**	2.791**
EGARCHN									-	-7.524***	-6.806**	-5.037***	-5.122***	-4.920***
EGARCHT										-	3.625**	-3.403**	-4.407***	-4.051***
EGARCHG											-	-3.666**	-4.489***	-4.149***
GJRGARCHN												-	-3.152**	-2.559**
GJRGARCHT													-	0.917
GJRGARCHG														-

A regression is run with the difference between column model and row model as dependent variable and a constant as independent variable. T-statistics are reported. Negative values of the Diebold–Mariano test show that the squared errors of the model listed first (column models) are lower than those of the model listed last (row models). ARCH and GARCH under the t-distribution are omitted due to very large MSEs. The significance levels are indicated by *, ** and ***, and correspond to a significance level of 10%, 5% and 1% respectively, using a two-tailed test. Each model produces 3,332 forecasts.

Table C4: Diebold Mariano (1995) test statistics of the S&P500 comparing all models at the one year horizon.

	SMA 6	SMA 12	EWMA	ARMA	ARCHN	ARCHG	GARCHN	GARCHG	EGARCHN	EGARCHT	EGARCHG	GJRGARCHN	GJRGARCHT	GJRGARCHG
SMA 6	-	1.320	-1.202	-1.068	1.312	1.344	0.316	-3.322***	2.422**	2.240**	2.269**	1.994*	1.168	1.510
SMA 12		-	-1.647	-2.5**	0.659	0.769	-1.075	-3.433***	3.086***	2.687**	2.752**	2.157**	0.481	1.120
EWMA			-	0.419	1.724	1.745	1.448	-4.168***	2.331**	2.239**	2.255**	2.116*	1.684	1.863
ARMA(1,1)				-	3.294***	3.372***	1.955	-2.912**	4.290***	4.058***	4.097***	3.775***	2.826**	3.204***
ARCH(N					-	2.299**	-2.127*	-3.464***	3.927***	3.223***	3.339***	2.460**	-0.060	0.867
ARCHGED						-	-2.224*	-3.471***	3.707***	3.017**	3.131***	2.265**	-0.224	0.692
GARCHN							-	-3.574***	3.979***	3.609***	3.672***	3.179***	1.662	2.256
GARCHGED								-	3.722***	3.679***	3.687***	3.624***	3.430***	3.518***
EGARCHN									-	-8.644***	-7.882***	-4.932***	-5.555***	-5.047***
EGARCHT										-	3.6649***	-3.112***	-4.834***	-4.168***
EGARCHG											-	-3.554***	-4.961***	-4.330***
GJRGARCHN												-	-3.767***	-2.859***
GJRGARCHT													-	1.289
GJRGARCHG														-

A regression is run with the difference between column model and row model as dependent variable and a constant as independent variable. T-statistics are reported. Negative values of the Diebold–Mariano test show that the squared errors of the model listed first (column models) are lower than those of the model listed last (row models). ARCH and GARCH under the t-distribution are omitted due to very large MSEs. The significance levels are indicated by *, ** and ***, and correspond to a significance level of 10%, 5% and 1% respectively, using a two-tailed test. Each model produces 3,101 forecasts.

Table C5: Diebold Mariano (1995) test statistics of the S&P500 comparing all models at the two year horizon.

	SMA 6	SMA 12	EWMA	ARMA	ARCHN	ARCHG	GARCHN	GARCHG	EGARCHN	EGARCHT	EGARCHG	GJRGARCHN	GJRGARCHT	GJRGARCHG
SMA 6	-	3.04***	-1.399	0.058	2.495**	2.494**	1.308	-4.201***	3.332***	3.228***	3.2511***	2.943***	2.121*	2.567**
SMA 12		-	-2.40**	-2.01*	0.649	0.752	-1.231	-4.447***	2.457**	2.255**	2.298**	1.736	0.236	0.959
EWMA			-	1.292	2.424**	2.436**	1.950	-4.579***	2.917***	2.867***	2.878***	2.726**	2.321**	2.536**
ARMA(1,1)				-	2.753**	2.822***	1.411	-4.125***	3.291***	3.185***	3.206***	2.883***	2.090*	2.484**
ARCH(N					-	2.014*	-3.224***	-4.471***	2.729**	2.433**	2.492**	1.671	-0.369	0.605
ARCHGED						-	-3.407***	-4.475***	2.580**	2.286**	2.344**	1.526	-0.511	0.461
GARCHN							-	-4.404***	4.389***	4.147***	4.195***	3.576***	1.862	2.682**
GARCHGED								-	4.588***	4.571***	4.575***	4.526***	4.406***	4.475***
EGARCHN									-	-8.538***	-7.647***	-5.051***	-5.102***	-4.484***
EGARCHT										-	2.964***	-3.726***	-4.674***	-3.910***
EGARCHG											-	-4.0192***	-4.766***	-4.035***
GJRGARCHN												-	-3.402***	-2.216*
GJRGARCHT													-	1.501
GJRGARCHG														-

A regression is run with the difference between column model and row model as dependent variable and a constant as independent variable. T-statistics are reported. Negative values of the Diebold–Mariano test show that the squared errors of the model listed first (column models) are lower than those of the model listed last (row models). ARCH and GARCH under the t-distribution are omitted due to very large MSEs. The significance levels are indicated by *, ** and ***, and correspond to a significance level of 10%, 5% and 1% respectively, using a two-tailed test. Each model produces 2,849 forecasts.