International Journal of Banking and Finance, Volume 9 (Number 1), 2012: pages 48-80

MODELLING AND FORECASTING VOLATILITY IN THE GOLD MARKET

Stefan Trück and Kevin Liang
Macquarie University, Australia

Abstract

We investigate the volatility dynamics of gold markets. While there are a number of recent studies examining volatility and Value-at-Risk (VaR) measures in financial and commodity markets, none of them focuses on the gold market. We use a large number of statistical models to model and then forecast daily volatility and VaR. Both in-sample and out-of-sample forecasts are evaluated using appropriate evaluation measures. For in-sample forecasting, the class of TARCH models provide the best results. For out-of-sample forecasting, the results were not that clear-cut and the order and specification of the models were found to be an important factor in determining model's performance. VaR for traders with long and short positions were evaluated by comparing failure rates and a simple AR as well as a TARCH model perform best for the considered back-testing period. Overall, most models outperform a benchmark random walk model, while none of the considered models performed significantly better than the rest with respect to all adopted criteria.

Key Words: Gold Markets, Volatility, Forecasting, Value-at-Risk, Backtesting
JEL classification: G17, C22, G32

1. Introduction

The recent global financial crisis has highlighted the need for financial institutions to find and implement appropriate models for risk quantification. Hereby, in particular Value-at-Risk (VaR) and volatility estimates were subject to significant changes during 2007-9 financial turmoil in comparison to normal market behaviour. Further, as the risk in equity and bond markets was increasing, there was a particular interest of investors to increase their positions in the gold market. This study evaluates the

effectiveness of various volatility models with respect to forecasting market risk in the gold bullion market. While there is a stream of literature examining performance of models for volatility and VaR, this is a pioneer study to particularly focus on the gold market. Despite the important role gold plays for risk management and hedging in financial markets, there has been relatively little literature on the estimation of volatility of gold. Exceptions include the studies by Mills (2003), Tully and Lucey (2006), Canarella and Pollard (2008), Morales (2008) and Jun (2009).

Generally, the gold market has a significant and unique role in financial markets as a safe haven that is also used for hedging and diversification. While there is no theoretical reason why gold is referred to as a safe haven asset, historical evidence suggests that investments in the gold market spikes during times of turmoil in other financial markets. One explanation could be that it is one of the oldest forms of money and was traditionally used as an inflation hedge. Moreover, gold is often uncorrelated or even negatively correlated with other types of assets. This is an important quality that allows gold to act as a diversification asset in portfolios, since a more globalised market has led to the increase in correlation among other assets. This became also evident during the financial crisis of 2007-2009 where the negative effect of one market readily flowed into other markets, yet the gold market remained relatively unscathed during this period of turbulence. So far there has been no study using volatility and VaR modelling in the spot gold and gold futures markets. Gold market research has concentrated on the role of gold as a hedging or diversification tool, in particular as a safe haven during market crashes.

This study examines various models that can be used in forecasting volatility, to evaluate their respective performance. Finding appropriate models for volatility is of interest for several reasons: firstly, it is an integral factor of derivative security pricing, for example, in the classic Black-Scholes model or alternative option pricing formulas. Secondly, as a representation of risk, volatility plays an important role in an investor's decision making process. Volatility is not only of great concern for investors but also policy makers and regulators who are interested in the effect of volatility on the stability of financial markets in particular and the whole economy in general. Finally, volatility estimation is an essential input in many VaR models, as well as for a number of applications in a firms market risk management practices.

The remainder of the paper is set up as follows. Section 2 provides a brief review on the global gold market and studies on volatility modelling of financial markets in general and gold markets in particular. Section 3 provides an overview on the data and techniques used in this study. In particular various models for volatility forecasting and evaluating model performance are reviewed. Empirical results of the study are reported in Section 4 while section 5 concludes.

2. Gold Markets and Volatility Models

2.1 The gold market

Gold has been used throughout history as a form of payment and has been a standard for currency equivalents to many economic regions or countries. In spite of its historical monetary significance, a free functioning world market only came of age in recent times. Before 1971, the gold standard was mostly used in various times in history, where domestic currencies have been backed by gold. The system existed until 1971, when the US stopped the direct convertibility of the United States dollar to gold, effectively causing the system to break down. Since then, a global market for gold in its own right developed, remaining open around the clock and open to a range of derivative instruments.

The market for gold consists of a physical market in which gold bullions and coins are bought and sold and there is a paper gold market, which involves trading in claims to physical stock rather than the stock themselves. Physical gold is generally traded in the form of bullions. The bullion market serves as a conduit between larger gold suppliers such as producers, refiners and central banks and smaller investors and fabricators. The bullion market is essentially a spot market, but is complemented by the use of forward trading for the hedging of physical positions.

Since 1919, the most widely accepted benchmark for the price of gold is known as the London gold fixing, a twice-daily (telephone) meeting of representatives from

five bullion-trading firms.[1] Furthermore, there is active gold trading based on the intra-day spot price, derived from gold-trading markets around the world as they open and close throughout the day. The key prices in the London bullion market are the spot (fixings) price, the forward price and the lease rate. The spot (fixings) price is a daily clearing or *fix* price obtained by balancing purchases and sales ordered through its members. The forward price (GOFO) is the simultaneous purchase and sales price of gold forward contracts of various lengths. Generally, the GOFO rate is expressed as an annual percentage. Finally, the lease rate refers to short-term loans denominated in gold and is expressed as an annualized interest rate.

Since 1971 the price of gold has been highly volatile, ranging from a high of US$850 on January 21, 1980, to a low of US$252.90 on June 21, 1999. The period from 1999 to 2001 marked the so-called *Brown Bottom* after a 20-year bear market. Prices increased rapidly from 1991, but the 1980 high was not exceeded until 2008 when a new maximum of $865.35 was set on January 3, 2008. Another record price was set on March 17, 2008 at $1023.50. In the second half of 2009, gold markets experience renewed momentum upwards due to increased demand and a weakening US dollar. Overall, since April 2001, the gold price has more than tripled in value against the US dollar.

2.2 Factors influencing gold prices

As mentioned above, gold has a unique place in financial markets. Of all the precious metals, gold is the most popular as an investment. Investors generally buy gold as a hedge or safe haven against any economic, political, social or currency-based crises. These crises include investment market declines, burgeoning national debt, currency failure, inflation but also scenarios like war or social unrest. As in any commodities, the price of gold is ultimately driven by its supply and demand. However, unlike other resources, hoarding and disposal plays a much bigger role in price formation because most of the gold ever mined still exists and is potentially able to enter the market for the right price. Given the huge quantity of stored gold,

[1] All five members - Bank of Nova Scotia–ScotiaMocatta, Barclays Bank Plc, Deutsche Bank AG, HSBC Bank USA, NA and Société Générale are market making members of the London Bullion Market Association (LBMA).

compared to the annual production, the price of gold is mainly affected by changes in sentiment, rather than changes in the actual annual production.

Also macroeconomic factors such as low real interest rates can have an effect on gold price. If the return on bonds, equities and real estate is not adequately compensating for risk and inflation, then the demand for gold and other alternative investments such as commodities increases. An example of this is the period of stagflation that occurred during the 1970s which led to an economic bubble forming in precious metals.

Financial market declines such as the 2007-9 global financial crisis usually leads investors to look for alternative and less volatile investment opportunities for their funds. It will also increase the need for investors to hedge their portfolios to minimise their risk in case of further decline. The demand for gold and, thus, its price increase, empirically is due to the role of gold as a safe haven in times of crises. This is one of the major reasons to drive gold prices to new highs throughout the post-financial crisis period.

Central banks and the International Monetary Fund (IMF) also play an important role in determining the gold price. At the end of 2004 central banks and official organizations held 19 percent of all above-ground gold as official gold reserves. Thus, they have a significant influence on the gold market not only as a major buyer and seller. Also, speculation on their future gold holding levels can also be a driving factor. Recently, the assumption that central banks around the world will increase their gold reserve levels as a hedge against the falling US dollar has also contributed to the rise of gold prices.

The performance of gold bullion is often compared to stocks. However, they are fundamentally different asset classes. Gold is regarded by some as a store of value (without growth) whereas stocks are regarded as a return on value. Stocks and bonds perform best in periods of economic stability and growth, whereas gold is seen as the asset to hold in times of uncertainty and crisis. Throughout history there has been a cyclical run with long periods of stock outperformance followed by long periods of gold outperformance. Over the long term, equity markets have been able to outperform gold overall.

2.3 Volatility Models

Within the last three decades various approaches to volatility modelling have been suggested in the econometric and financial literature. In the following we will provide a brief overview of developments in the literature starting with the autoregressive conditional heteroskedasticity (ARCH) models (Engle, 1982). Bollerslev (1986) introduced the generalised ARCH (GARCH) model. The latter is often utilised in financial market studies. The general idea is to predict the current period's variance by forming a weighted average of a long term average, the forecasted variance from last period, and information about volatility observed in the previous period. If the return is unexpectedly large either in the upward or the downward direction, then the trader will increase the estimate of the variance for the next period. This model is also consistent with the volatility clustering often seen in financial returns data, where large changes in returns are likely to be followed by further large changes.

Since the introduction of these models, they have been widely used in volatility modelling and forecasting. Researchers such as French *et al.* (1987) and Akgiray (1989) utilised GARCH models to capture the behaviour of stock market price volatilities. Argiray (1989) compared the GARCH (1,1) model to other historical estimation methods and found that the GARCH (1,1) model outperformed its competitors. Many extensions of the GARCH model have been introduced in the literature since: e.g. GARCH-in-mean (GARCH-M) models (Engle *et al.*, 1987), EGARCH models (Nelson, 1991), Threshold ARCH (TARCH) and Threshold GARCH (TGARCH) (Glosten, Jaganathan, and Runkle, 1993; Zakoïan, 1994) and Power Arch (PARCH) models (Ding *et al*., 1993) just to name a few.

A number of studies have focused on optimal model specification and the performance of various GARCH models in financial markets providing no clear-cut results. Hansen and Lunde (2005) carried out comprehensive testing of 330 variants of ARCH type models on their performance in estimating volatility in exchange rates and stock returns. The study found that the GARCH (1,1) model outperforms other models in estimating exchange rate volatilities but underperforms in estimating stock returns. McMillan *et al*. (2000) tested a set of ten volatility estimation models including

random walk, moving average and GARCH models in forecasting UK stock market returns at different frequencies. They found that the performance of each model varied depending on the length of frequencies, the series as well as the type of loss function being applied. The random walk model outperformed others at the monthly frequency, while GARCH and moving average models were superior using daily forecasts. Brooks and Persand (2002, 2003) examine various ARCH and GARCH type models with respect to volatility forecasting. They report that, while the forecasting performance of the models depended on the considered data series and time horizon, the overall most preferred model is a simple GARCH(1,1). This is also consistent with many other studies such as e.g. Bollerslev *et al*. (1992). On the other hand, Braisfold and Faff (1996) evaluate volatility models in forecasting stock returns, and find that none of the models significantly outperforms the others.

Recently, also a stream of literature has emerged focusing on modelling and forecasting volatility with respect to the quantification of Value-at-Risk (VaR). As pointed out by Jorion (1996), VaR plays a substantial role in managing risks for financial institutions. The importance of the VaR measure is further highlighted by regulators in the Basel Committee on Banking Supervision.[2] The performance of volatility models with respect to appropriate quantification of VaR has been investigated by Danielsson and De Vries (2000): conditional parametric methods such as the GARCH model significantly underpredict the VaR of U.S. stock returns. Giot and Laurent (2001, 2003) investigate volatility models for both negative and positive returns, with the latter representing risk for short position holders. They find that skewed asymmetric ARCH models using the Student t distribution perform best with respect to risk quantification. Sadorsky (2006), investigating oil price volatility, tested a great variety of volatility models by evaluating the forecasting performance using different VaR measures. His findings suggest that while no model could consistently outperform the others, a GARCH model as well as a TGARCH performed quite well for modelling and forecasting the volatility and risk of oil prices.

Tully and Lucey (2006) examine various macroeconomic influences on gold using models including the asymmetric power GARCH model (APGARCH) for spot

[2] The Basle Committee on Banking Supervision announced in April 1995 that capital adequacy requirements for commercial banks are to be based on VAR.

and futures prices over a 20 year period, paying special attention to periods of stock market crashes. Their results suggest that the price of gold is significantly influenced by the U.S. dollar while during periods of financial crises an APGARCH model performs best with respect to volatility. Mills (2003) investigates the statistical behaviour of daily gold prices, and finds that price volatility scaling with long-run correlations is important while gold returns are characterised by short-run persistence and scaling with a break point of 15 days. Canarella and Pollard (2008) apply power GARCH model to the London Gold Market Fixings to investigate long memory features as well as conditional volatility behaviour of the returns. They find that APGARCH models were able to adequately capture long memory in returns and that market shocks have strong asymmetric effects: conditional volatilities of gold prices are affected more by good news (positive shocks) than bad news (negative shocks). Morales (2008) discusses volatility spill-over effects between precious metal markets using GARCH and EGARCH techniques. Gold was found to be influenced by prices of other precious metals, but there was little evidence to suggest other precious metals influencing gold prices.

3. Data and Models

3.1 The Data

The data for this study are daily PM gold fixing prices on the London Bullion Market available from the official The London Bullion Market Association website (www.lbma.org.uk). The market is a wholesale over-the-counter (OTC) market for gold and silver. The fixings are the internationally published benchmarks for precious metals. The Gold Fixing is conducted twice a day by five Gold Fixing members, at 10:30 am and 3:00 pm. This study will use the daily PM fixings price released at 3:00 pm as quoted in USD. The data cover 2508 observations from 4 January 1999 to 30 December 2008. The time series exhibits a number of price shocks, e.g. during the periods around September 11, 2001, the beginning of the Iraq war in 2003 as well as the global financial crisis in 2008.

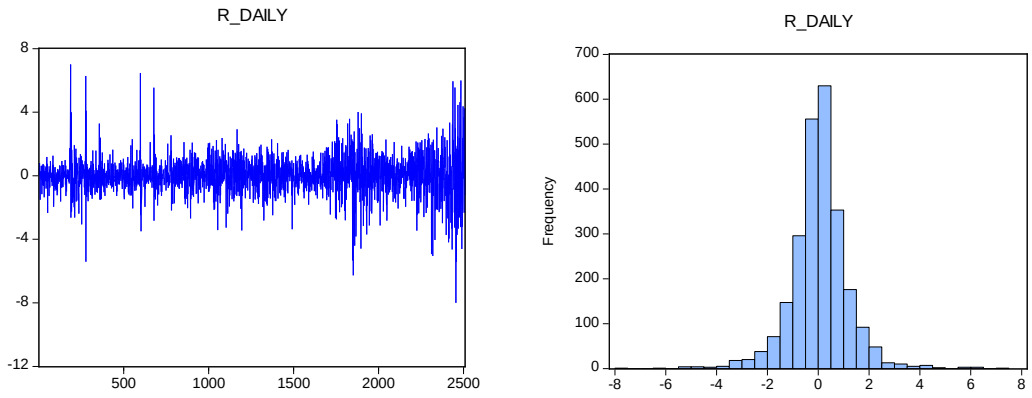
For the observed gold fixing prices p_t, the daily log-returns are calculated as $r_t = \ln(p_t/p_{t-1})$. Table 1 provides a summary of descriptive statistics for the considered return series.

Table 1: Summery Statistics of Gold-Fixing Log Returns

	Gold Fixing
Mean	0.044204
Median	0.045767
Maximum	7.005954
Minimum	-7.971887
Std. Dev.	1.143359
Skewness	-0.053361
Kurtosis	8.533989
Jarque-Bera	3200.230
Probability	0.000000
Sum	110.8201
Sum Sq. Dev.	3276.017
Observations	N = 2507

We observe that the mean and median of daily returns are positive indicating that overall gold prices were increasing during the considered time period. The magnitude of the average return (0.044%) is very small in comparison to its standard deviation (1.14%). Further, the large kurtosis of 8.53 indicates the leptokurtic characteristics of daily returns.

Figure 1: Time series and distribution of returns for gold fixing prices, 1999 to 2008



Obviously, the series has a distribution with tails that are significantly fatter than those of a normal distribution. This indication of non-normality is also supported by the Jarque and Bera (1980) test statistic, which rejects the null hypothesis of a normal distribution at all levels of significance. Figure 1 provides a plot of the time series for the daily log-returns as well as a histogram of the return distribution. The figures indicate heteroscedasticity and volatility clustering for the return series that also exhibits a number of rather isolated extreme returns caused by unforeseen events or shocks to the gold market. We further test for stationarity of the return series using the Augmented Dick Fuller (1979) (ADF) and Phillips Perron (1988) (PP) unit root tests. The ADF test is set to a lag length 0 using the Schwarz Information Criterion (SIC) and the PP test is conducted using the Bartlett Kernel spectral estimation method. Results are reported in Table 2, and indicate that for both tests the null hypothesis of a unit root is rejected. So the return series gold fixing prices can be considered to be stationary.

Table 2: Results for Augmented Dickey-Fuller and Phillips-Perron unit root tests for gold fixing return series (4 January 1999 -30 December 2008)

	t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic	-49.24794	0.0001
	Adj. t-Stat	**Prob.***
Phillips-Perron test statistic	-49.29521	0.0001

3.2 Considered Models

In the following, a variety of models is introduced for volatility modelling and forecasting of the daily returns. We will follow several studies in the literature, see e.g. Sadorsky (2006), and measure the volatility of gold by its squared daily return:

$$\hat{\sigma}_t^2 = r_t^2 \qquad (1)$$

Thus, most of the models will be evaluated with respect to their ability to model and forecast the volatility measured by the squared return of the gold fixings price. The first model to be considered in the empirical analysis is a random walk model

(RW). If the volatility of gold market returns follows a random walk, the best forecast for the next period's volatility is the volatility observed in the current period:

$$\hat{\sigma}_{t+1}^2 = r_t^2 \qquad (2)$$

This random walk model will be used as a benchmark model for the out-of-sample performance of the estimated models.

The second standard class of models to be considered are historical mean (HM) models. In these models, the forecast for the volatility of the next period is the average of all previous volatilities. In particular, if $\hat{\sigma}_t^2$ is a random variable, which is uncorrelated with other observable variables and if $\hat{\sigma}_t^2$ is uncorrelated with its own past values, then the population mean can be considered as the optimal forecast. Defining $\sigma_t^2 = r_t^2$, the HM model can be denoted by

$$\hat{\sigma}_{t+1}^2 = \frac{1}{T}\sum_{i=0}^{T-1}\sigma_{t-i}^2 \qquad (3)$$

A popular alternative to the HM model is the m-period moving average (MA) model. The forecast for the next period is based on the average of the last m observations. A value for m has to be determined. We decided to use moving averages of length m=20, 40 and 120 days, corresponding to about one month, two months and six months. The MA(m) model can be denoted by:.

$$\hat{\sigma}_{t+1}^2 = \frac{1}{m}\sum_{i=0}^{m-1}\sigma_{t-i}^2 \qquad (4)$$

The next model we consider is the exponentially weighted moving average model (EWMA). It forecasts the future volatility by applying weighting factors which decrease exponentially. That is, the method gives higher weights to more recent observations while still not discarding older observations entirely. It is calculated as the weighted average of the estimated volatility $\hat{\sigma}_t^2$ for day t (made at the end of day t-1) and the value of volatility σ_t^2 observed on day t:

$$\hat{\sigma}_{t+1}^2 = \alpha\hat{\sigma}_t^2 + (1-\alpha)\sigma_t^2 \qquad (5)$$

The smoothing parameter α governs how responsive the forecast is to the most recent daily percentage change. Generally, α lies between 0 and 1, and the process becomes a RW for α =0. A popular choice for the parameter α is based on J.P.

Morgan's RiskMetrics (1995) where it is suggested that $\alpha = 0.94$ provides forecasts of the variance rate closest to the actual variance rate for a range of different market variables.

An alternative is an ordinary least squares (OLS) model. The relationship between volatility on day *t* and day *t+1* is described based on a linear relationship:

$$\hat{\sigma}_{t+1}^2 = \hat{\beta}_0 + \hat{\beta}_1 \hat{\sigma}_t^2 \qquad (6)$$

The parameter estimates are then determined by OLS estimation. The model can be extended to an autoregressive (AR) model of order p where the current volatility is a linear function of the last p observations for the volatility. We implement a model of order p = 5 such that we estimate an AR(5) model that can be described by the following equation:

$$\hat{\sigma}_{t+1}^2 = \hat{\beta}_0 + \hat{\beta}_1 \hat{\sigma}_t^2 + \hat{\beta}_2 \hat{\sigma}_{t-1}^2 + \hat{\beta}_3 \hat{\sigma}_{t-2}^2 + \hat{\beta}_4 \hat{\sigma}_{t-3}^2 + \hat{\beta}_5 \hat{\sigma}_{t-4}^2 \qquad (7)$$

We also consider a weighted moving average of disturbance terms model (MAD) where the volatility in period t+1 is modelled as a function of the lagged values of the disturbance term ε_t. Similar to the AR model, we decided to use a MAD model of order 5 that can be described by the following equation:

$$\hat{\sigma}_{t+1}^2 = \hat{\alpha}_0 + \hat{\alpha}_1 \varepsilon_t + \hat{\alpha}_2 \varepsilon_{t-1} + \hat{\alpha}_3 \varepsilon_{t-2} + \hat{\alpha}_4 \varepsilon_{t-3} + \hat{\alpha}_5 \varepsilon_{t-4} \qquad (8)$$

We decided to also use an autoregressive moving average (ARMA) or Box-Jenkins model that includes both an autoregressive (AR) and a moving average (MAD) component. A simple ARMA(1,1) can then be described by the following equation:

$$\hat{\sigma}_{t+1}^2 = \hat{\beta}_0 + \hat{\beta}_1 \hat{\sigma}_t^2 - \hat{\alpha}_1 \varepsilon_t \qquad (9)$$

Since the introduction of autoregressive conditional heteroscedasticity (ARCH) models by Engle (1982), the ARCH and even more the related GARCH (Bollerslev, 1986) model have become standard tools for examining the volatility of financial variables. The model has proven to be very useful in capturing heteroskedastic behaviour or volatility clustering without the requirement of higher order models in various financial markets, see e.g. Choudhy (1996) or Sadorsky (2006). In a GARCH (1,1) model the conditional variance equation can be denoted by

$$h_t = \omega + \alpha\varepsilon_{t-1}^2 + \beta h_{t-1} \quad (10)$$

while the equation for the conditional mean is

$$r_t = \pi + \varepsilon_t,\ \varepsilon_t \sim N(o, h_t) \quad (11)$$

such that the one day ahead variance forecast can be expressed as:

$$\hat{h}_{t+1} = \hat{\omega} + \hat{\alpha}\hat{\varepsilon}_t^2 + \hat{\beta}\hat{h}_t \quad (12)$$

A popular extension of the GARCH (1,1) model is also the GARCH in mean (GARCH-M) model that was first proposed by Engle *et al*. (1987). The GARCH-M model includes the conditional variance in the specified equation for the conditional mean. This allows for so-called time varying risk premiums. Chou (1988) suggests that the dynamic structure of the conditional variance can be captured more flexibly by a GARCH-M model, using the following specification for the conditional mean:

$$r_t = \pi + \delta h_t + \varepsilon_t,\ \varepsilon_t \sim N(o, h_t) \quad (13)$$

Another extension of standard ARCH and GARCH models has been suggested by Glosten *et al*. (1994) and Hentschel (1994): threshold ARCH (TARCH) and GARCH (TGARCH) models, which are popular in describing return asymmetry. Large negative returns are often followed by a substantial increase in volatility such that the TARCH and TGARCH models distinguish between negative and positive returns. The TGARCH model that will be considered in the empirical analysis treats the conditional standard deviation as a linear function of shocks and lagged standard deviations (Hentschel, 1994) and is denoted by:

$$h_t = \omega + \alpha\varepsilon_{t-1}^2 + \beta h_{t-1} + \gamma\varepsilon_{t-1}^2 D_{t-1} \quad (14)$$

where D_{t-1} is equal to 1 if $\varepsilon_t < 0$, and zero otherwise. Obviously, in this model, $\varepsilon_{t-1}^2 > 0$, and $\varepsilon_{t-1}^2 < 0$ will have different effects on the conditional variance. If $\gamma \neq 0$, there is asymmetry in the model. If $\gamma > 0$, the occurrence of bad news will increase volatility and there is evidence of a leverage effect.

3.3 Performance Evaluation Measures

To evaluate the performance of the considered models, we apply a variety of measures such as mean squared error (MSE), root mean squared error (RMSE), mean absolute deviation (MAD), mean absolute percentage error (MAPE) and the Theil U statistic. The MSE quantifies the difference between predicted and actually observed values by considering the squared difference between these two quantities:

$$MSE = \frac{1}{t}\sum_{t=1}^{T}(\sigma_t^2 - \hat{\sigma}_t^2)^2 \tag{15}$$

The RMSE is simply the root of MSE and has the advantage of being measured in the same unit as the forecasted variable:

$$RMSE = \sqrt{\frac{1}{t}\sum_{t=1}^{T}(\sigma_t^2 - \hat{\sigma}_t^2)^2} \tag{16}$$

The MAE is also measured in the same unit as the forecast, but gives less weight to large forecast errors than the MSE and RMSE:

$$MAE = \frac{1}{t}\sum_{t=1}^{T}|\sigma_t^2 - \hat{\sigma}_t^2| \tag{17}$$

The MAPE measures the forecast quality independent of the unit of measurement of the variable. The measure might be less useful when the actual values of σ_t^2 are close to zero, because in this case the MAPE will take on large values even if the errors are fairly small in magnitude. Another drawback of the MAPE is that if there are zero values (which may happen for daily squared returns) there will be a division by zero. We still decided to examine the results using the MAPE measure that can be denoted by:

$$MAPE = \frac{100}{t}\sum_{t=1}^{T}\frac{|\sigma_t^2 - \hat{\sigma}_t^2|}{\sigma_t^2} \tag{18}$$

We also investigate the forecasting performance using the Theil U statistic that examines the RMSE measure of a forecast against a naïve one step ahead forecast. If

the Theil U statistic is smaller than 1, the tested forecast model outperforms the naïve model: if the U statistic is larger than 1, the naïve forecast is the better model. Note that in our analysis we decided to use the RW model as the naïve benchmark model for forecasting.

$$U = \frac{RMSE(forecast)}{RMSE(naiveforecast)} \qquad (19)$$

While the above forecasting quality measures are useful for providing different performance measures on applied models, they do not statistically test if the models are significantly different or better from another. Therefore, we will also apply the Diebold-Mariano (1995) test (DM) to compare the predictive ability between two forecasting models. The null hypothesis of the test is that the predictive ability of two forecasting models is the same. In our empirical analysis, we are particularly interested whether our forecast models are able to significantly outperform a simple RW model such that the considered models are tested against the RW model using a simple t-test, see e.g. Diebold (1998). Thus, the null hypothesis of equal performance of the models is rejected when the test-statistic

$$D = \frac{\bar{d}}{\hat{\sigma}/\sqrt{n}} \qquad (20)$$

with

$$d = \frac{1}{n}\sum_{t=1}^{n}(e_{1t}^2 - e_{2t}^2) \qquad (21)$$

yields significant values. In the empirical analysis we will restrict ourselves to one-period-ahead forecasts only. Note that the test could also be applied to k-step-ahead forecasts, see e.g. Diebold and Mariano (1995). The authors point out that the test tends to be less accurate for small sample sizes and k-step-ahead forecasts. However, these issues are unlikely to affect our empirical analysis due to a comparably large sample size and the use of one-period-ahead forecasts only.

4. Empirical Results

4.1. In-sample forecasting performance

In this section, we compute the one-step-ahead volatility forecasts using the models described in the previous section. For the in-sample analysis, the data are divided into three sub-periods: sub-period 1 from 28th Jun 1999 - Dec 2004, which is a period of slightly increasing gold prices over 6 years; sub-period 2 from Jan 2005 - Dec 2007, which is a period of substantially increasing gold prices over 3 years; and sub-period 3 from Jan 2008 - Dec 2008, a period of very volatile gold prices.

For in-sample forecasting, all observations within the period are used to estimate the models, and the forecasting results are compared to the actual values. The complete results including the ranking for each measure in each sub-period are presented in tables 3, 4 and 5, respectively. Obviously, the forecast error statistics, MSE, RMSE and MAE depend on the scale of the dependent variable and the differences between the actual and forecasted volatility. On the other hand, the MAPE does not depend on the scale of the variable: given a perfect fit of a model, the MAPE would be zero while there is no upper bound for the MAPE. For all measures, it can be concluded that the smaller the error statistic, the better the forecasting ability of a model. In order to facilitate the comparison of the models, for each performance criteria also the relative ranks are provided.

In the first sub-period, the price fluctuations were relatively low with a general upward trend. Only one structural break occurred after the 11 September 2001 attack. The shock lasted only a short period of time and did not have long-lasting effects on the volatility. The results show that for the MSE, MAE, RMSE and Theil U criteria the TARCH model yields the best results with respect to in-sample one period ahead forecasts. The model further ranks second for the MAPE criterion. Overall, most test statistics are consistent in ranking the forecasting performance of the considered models, with the exception of the MAPE measure. However, as discussed earlier, the MAPE can be unreliable in cases where the denominator contains the value of zero as it was the case in our evaluation where results were removed when the daily change and volatility were zero.

Table 3: In-sample Results for Sub-period 1

	RW	HM	MA(20)	MA(40)	MA(120)	OLS	AR(5)	MAD(5)	ARMA(1,1)	EWMA	GARCH(1,1)	GARCH-M	TARCH
MSE	10.231	7.617	7.937	7.798	7.615	6.774	6.708	6.718	7.615	7.472	7.907	7.258	5.713
Rank	13	10	12	11	9	4	2	3	8	6	11	5	1
RMSE	3.199	2.760	2.817	2.793	2.760	2.603	2.590	2.592	2.760	2.733	2.812	2.694	2.390
Rank	13	10	12	11	9	4	2	3	8	6	11	5	1
MAE	1.209	1.054	1.064	1.058	1.066	1.029	1.014	1.018	1.065	1.032	1.100	1.064	0.990
Rank	13	6	8	7	9	4	2	3	9	5	12	8	1
MAPE	8264.1	11071.5	8413.4	10260.2	10723.3	10263.7	9707.6	9913.2	10722.3	9114.0	9892.5	9904.0	8288.1
Rank	1	13	3	9	12	10	5	8	11	4	6	7	2
Theil U	1.000	0.863	0.881	0.873	0.863	0.814	0.810	0.810	0.863	0.855	0.879	0.842	0.747
Rank	13	9	12	10	8	4	2	3	7	6	11	5	1
DM	t-stat	-1.217	-1.082	-1.140	-1.228	-1.893	-1.912	-1.903	-1.228	-1.313	-1.082	-1.457	-2.348
Rank		9	12	10	8	4	2	3	7	6	11	5	1
	p-value	0.112	0.140	0.127	0.110	0.029	0.028	0.029	0.110	0.095	0.140	0.073	0.010

Table 3 In-sample forecast results for sub-period 1 (June 1999 - December 2004) for examined models and the considered performance evaluation measures MSE, RMSE, MAE, MAPE, Theil U as well as results for Diebold-Mariano test with RW model as benchmark.

Table 4: In-sample Results for Sub-period 2

	RW	HM	MA(20)	MA(40)	MA(120)	OLS	AR(5)	MAD(5)	ARMA(1,1)	EWMA	GARCH(1,1)	GARCH-M	TARCH
MSE	15.203	8.347	7.704	7.706	7.829	8.153	7.595	7.762	7.558	7.607	7.558	7.682	6.624
Rank	13	12	5	6	8	9	3	7	2	4	2	6	1
RMSE	3.899	2.889	2.776	2.776	2.798	2.855	2.756	2.786	2.749	2.758	2.748	2.772	2.574
Rank	13	12	5	6	8	9	3	7	2	4	2	6	1
MAE	1.785	1.325	1.426	1.422	1.424	1.487	1.467	1.481	1.418	1.421	1.434	1.443	1.314
Rank	13	4	9	7	8	12	10	11	5	6	8	9	1
MAPE	14790.9	16439.6	15475.7	15515.4	19616.2	22234.2	18682.8	19663.1	17411.6	16083.5	17535.2	17679.1	9491.3
Rank	4	8	5	6	11	13	10	12	9	7	8	9	1
Theil U	1.000	0.741	0.712	0.712	0.718	0.732	0.707	0.715	0.705	0.707	0.705	0.711	0.660
Rank	13	12	5	6	8	9	3	7	2	4	2	6	1
DM	t-stat	-2.819	-3.113	-3.093	-3.038	-2.932	-3.055	-3.033	-3.173	-3.162	-3.138	-3.184	-3.567
Rank		12	6	7	9	11	8	10	3	4	5	2	1
	p-value	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.001	0.002	0.001	0.000

Table 4 In-sample forecast results for sub-period 2 (January 2005 - December 2007) for examined models and the considered performance evaluation measures MSE, RMSE, MAE, MAPE, Theil U as well as results for Diebold-Mariano test with RW model as benchmark.

Table 5: In-sample Results for Sub-period 3

	RW	HM	MA(20)	MA(40)	MA(120)	OLS	AR(5)	MAD(5)	ARMA(1,1)	EWMA	GARCH(1,1)	GARCH-M	TARCH
MSE	102.833	61.154	50.452	49.695	51.953	53.201	48.378	48.168	49.789	50.017	49.822	51.133	44.288
Rank	13	12	8	4	10	11	3	2	5	7	6	9	1
RMSE	10.141	7.820	7.103	7.049	7.208	7.294	6.955	6.940	7.056	7.072	7.058	7.151	6.655
Rank	13	12	8	4	10	11	3	2	5	7	6	9	1
MAE	5.758	4.668	4.205	4.060	3.820	4.509	4.360	4.410	4.083	4.129	4.016	4.148	3.812
Rank	13	12	8	4	3	11	9	10	5	6	4	7	2
MAPE	11637.6	6670.9	11419.8	10527.5	8212.6	13501.3	12257.8	12197.5	10654.3	10654.9	9977.4	9912.7	10973.9
Rank	10	1	9	5	2	13	12	11	6	7	4	3	8
Theil U	1.000	0.771	0.700	0.695	0.711	0.719	0.686	0.684	0.696	0.697	0.696	0.705	0.656
Rank	13	12	8	4	10	11	3	2	5	7	6	9	1
DM	t-value	-9.769	-12.394	-12.498	-11.948	-11.765	-12.467	-12.699	-12.549	-12.534	-12.391	-12.475	-13.876
Rank		12	8	6	10	11	5	2	3	4	9	7	1
	p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 5 In-sample forecast results for sub-period 3 (January 2008 - December 2008) for examined models and the considered performance evaluation measures MSE, RMSE, MAE, MAPE, Theil U as well as results for Diebold-Mariano test with RW model as benchmark.

The different ranking according to the MAPE criterion is also indicated by the fact that while the benchmark RW model ranks last for all other statistics, it is the preferred model according to the MAPE. The performances of the other GARCH type models vary substantially depending on the considered evaluation criterion. Interestingly, the GARCH (1,1) model performs poorly and is ranked in the bottom quartile of all tested models. The GARCH-M model performs better, ranking 5th out of the 13 models in most statistics. The simple AR(5) and MAD(5) models also perform well, producing the second and third lowest values for the considered error statistics. Other ARMA type models including the OLS and ARMA(1,1) rank in the middle, while models based on past averages such as the HM, and the various MA models do not perform very well. However, they still outperform the RW model for most of the considered criteria.

The Theil U statistics are less than 1 for all considered models being indicative of the models performing better than the random walk model. On the other hand, the results for the DM test are not so clear-cut. Only the OLS, AR, MAD and TARCH models are significantly better than the benchmark random walk model at the 0.05 significance level. So during the period with low volatility from June 1999 and December 2004, all models seem to provide better in-sample forecasts than the RW model, while only 4 out of 12 significantly outperform the RW model according to the DM test. We will see that during the more volatile periods, the DM test provides more significant results. We could conclude that during periods of low volatility in the gold market, it seems difficult to significantly outperform the RW model using a parametric model. Overall, while the TARCH model had the lowest forecast errors of all models, one may also consider AR, MA and OLS type models when forecasting in periods of low volatility.

As mentioned above, during the second sub-period from January 2005 and Dcember 2007, the gold market experienced a major boom, with gold almost doubling in value. Overall, the forecast errors for this period are slightly higher than that of sub-period 1. Once again the TARCH model turns out to be the best model considering the different performance criteria. Also the ARMA(1,1) and GARCH(1,1) models provide good results and they rank among the best three models for several of the performance criteria. Interestingly, for this period, different performance measures provide quite

different rankings: Theil's U allocates more weight to large deviations while MAE and MAPE provide a different weighting of the performance errors. Thus, the conditional variance type models GARCH(1,1) and GARCH-M models were selected by MSE, RMSE, and Theil's U as the second and sixth best models. In fact, these perform worse against the absolute error measures of MAE and MAPE. The MSE and Theil's U also indicate that the EWMA, AR(5) and ARMA models perform well.

The RW is once again the worst performing model, ranking last for all statistics except MAPE. The DM values for this period are all highly significant even at the 0.01 level, indicating that most models are able to significantly outperform the RW benchmark in this period. This is also confirmed by U statistic where all models yield lower values than in the first sub-period. The U values range from 0.26 to 0.34 indicating that even the worst performing model (HM) is still significantly better than the RW benchmark. Overall, the results for the second sub-period suggest that predictive models with conditional volatility like TARCH, GARCH and GARCH-M seem to perform quite well during this period of significant increases in the gold price.

The third sub-period from January to December 2008 also includes the advent of the global financial crisis, when various financial markets as well as the gold market exhibited a long period of extreme volatility. Generally, one would expect this period being the most difficult for volatility prediction. This is confirmed by both MSE and MAE-based criteria yielding clearly higher values than for the previous two sub-periods. For example, the MSE is five times higher than during the first and second sub-period while the MAE increases by roughly 200 percent. Also for the third sub-period, MSE, RMSE and U favour the TARCH model as yielding the best predictions, while the AR(5) and MAD(5) rank second and third. For these criteria, the random walk model is the worst performing model, followed by the HM model. Also the MAE measure gives indication of superiority of the TARCH model over the others. However, for this criterion, the AR and MAD models perform rather poorly and only rank ninth and tenth. Again, the two worst performing models are the RW and HM model.

The DM test show that for the third sub-period all models were able to significantly outperform the RW model at the 0.01 level. Results for Theil's U are

similar to the second sub-period indicating that the models provide substantially smaller RMSE than the RW model for the volatile third sub-period. Overall, we conclude that for in-sample fit, the TARCH model can be considered as the most appropriate, ranking first for almost all of the examined performance measures and sub-periods.

4.2. Out-of-sample forecasting results

In the following we report the results for an out-of-sample analysis of the models by comparing one-step-ahead volatility for the most volatile period from July 1, 2008 to December 30, 2008. A recursive window approach is used. For the recursive window approach, the initial estimation date is fixed and the models are estimated using all observations available up to the initial estimation date. It is an iterative procedure, where in each time step, the estimation sample is augmented to include one additional observation in order to re-estimate the volatility forecast for the next day. Again, results are benchmarked against a RW model. Note that despite its simplicity, particularly in out-of-sample forecasting the random walk model is often considered as a benchmark model that is difficult to beat: for example, Stock and Watson (1998) examine various US macroeconomic time series and suggest the RW model to perform best amongst a number of competing models.

The out-of-sample results for the different models are provided in Table 6. Our results for the MSE criterion suggest that the MA(40) model provides the most accurate forecasts while the EWMA model ranks seconds. Interestingly, similar to the considered in-sample periods, the RW model proved to be the worst amongst the examined models also for out-of-sample forecasting. It ranked last with respect to the MSE criterion and provided predictions significantly less accurate than most of the considered models. Another feature of the results is that there are only relatively small differences with respect to MSE among the ten best models: the MSE for the MA(40) model is 83.29 while the MAD(5) model provides a MSE of 90.08.

With respect to MAE, we observe the smallest error for the MA(120) model. The HM and MAD models, also perform well, ranking second and third, respectively. The benchmark RW model is substantially less accurate than the other models. Again the marginal difference between the first and tenth ranked model is comparably small.

Overall, as could be expected, the out-of-sample errors are higher than the comparable numbers for the in-sample results. For MSE, there is an average increase by approximately 30 percent while for MAE, the average errors increase by approximately 80 percent in comparison to the in-sample results for the third sub-period. Again choosing the MAPE statistics provides slightly different findings and suggests the HM model as the best model followed by the OLS model. Once again the RW model performs worst.

According to the U statistic, all models performed better than the RW model. Since U is based on the MSE measure, the results suggest the MA(40) followed by the EWMA model as being most appropriate. Generally the differences between the results for all models are rather small, since the values for Theil's U range from 0.70 to 0.78. Results are confirmed by the DM test indicating that all models are significantly better than the benchmark RW model at the 0.01 level of significance.

Considering the different model types we find that MA models performed clearly better in out-of-sample than in-sample forecasting, in particular the MA(40) model. Yet again there was no clear-cut outperformance of this model class with the MA(20) ranking fourth and the MA(120) ranking eighth. The HM model was the second worst according to the MSE, but it is one of the better models according to the MAE and MAPE. For the GARCH models it is noteworthy that the TARCH model, which provided superior in-sample fit for all three sub-periods, yielded one of the worst results for out-of-sample forecasting. On the other hand, in terms of the magnitude of MSE and MAE, the difference with the best model was still rather small. The ARMA models rank second to eleventh across the different measures indicating the importance of the right choice of the order of the coefficients.

In summary, we conclude that there are only small differences with respect to the out-of-sample forecast performance between the considered models. The MA(40) could be considered the best model based on the MSE and U measures. Other models that have performed well are the ARMA(1,1) and the EWMA model. Furthermore, despite their generally good performance in the in-sample periods, for the considered out-of-sample period the GARCH models did not perform that well. In particular the TARCH model, that was the clear winner when in-sample volatility predictions were

considered, only ranked between 9 and 13 across the measures. Overall, there are no significant differences between the models and the rankings based on each performance measure are quite different.

We conclude that, for the out-of-sample forecasting, it is hard to choose an overall winner. We will now extend our analysis by examining the different models with respect to risk quantification. In particular, we investigate and report their performance in forecasting Value-at-Risk (VaR).

4.3 Value at risk Analysis

In this section, we examine the proposed models with respect to adequate VaR quantification in an out-of-sample forecasting study. For a given portfolio, probability and time horizon, VaR is defined as a threshold value of the probability that the mark-to-market loss of the portfolio over the given time horizon exceeds this value at a given probability level. In or analysis, following Giot and Laurent (2001, 2003), we evaluate the VaR forecasts from the perspective of both long and short position traders. The empirical models discussed in the previous sections are used to estimate and forecast the volatility $\hat{\sigma}_t$. The VaR for a given level of significance α can then be determined as VaR = $Z_\alpha \hat{\sigma}_t$. In our analysis we consider one day 95% and 99% VaR and, thus, set α equal to 0.05 and 0.01.

The corresponding failure rates and VaR violations are then computed by comparing the one period ahead forecasts of VaR with the actual observed returns in the out-of-sample period with 127 observations. We define the number of violations for long traders as being equal to the number of times the negative observed return on a particular day exceeds the one-day-ahead VaR forecasts. Correspondingly, the number of violations for a short position is the number of times the positive return is larger than the determined VaR forecast for that particular day. The failure rate is the percentage of violations occurring in the out-of-sample period. If the model has been specified correctly, the failure rate should approximately be equal to the theoretical number of exceptions at the chosen VaR level, see e.g. Kupiec (1995); Christoffersen (1998); Christoffersen and Diebold (2000) or Hull (2007). The results for the calculated VaR forecasts for long and short positions in the gold market are provided in Table 7 and 8.

Table 6: Out-of-sample Forecast Results for the Period July 1, 2008 to December 30, 2008

	RW	HM	MA(20)	MA(40)	MA(120)	OLS	AR(5)	MAD(5)	ARMA(1,1)	EMWA	GARCH(1,1)	GARCH-M	TARCH
MSE	172.870	105.245	84.649	83.291	87.972	101.224	86.081	90.080	84.249	83.812	86.942	86.352	88.238
Rank	13	12	4	1	8	11	5	10	3	2	7	6	9
RMSE	13.148	10.259	9.200	9.126	9.379	10.061	9.278	9.491	9.179	9.155	9.324	9.293	9.394
Rank	13	12	4	1	8	11	5	10	3	2	7	6	9
MAE	8.124	5.433	5.966	5.722	5.401	5.654	5.673	5.546	5.681	5.844	5.569	5.685	5.855
Rank	13	2	12	9	1	5	6	3	7	10	4	8	11
MAPE	10101.6	2165.3	10040.5	9429.3	7006.9	3936.8	6460.6	5060.6	8460.3	9513.5	7889.5	7747.1	13112.8
Rank	12	1	11	9	5	2	4	3	8	10	7	6	13
Theil U	1.000	0.780	0.700	0.694	0.713	0.765	0.706	0.722	0.698	0.696	0.709	0.707	0.714
Rank	13	12	4	1	8	11	5	10	3	2	7	6	9
DM	t-value	-1.764	-2.432	-2.457	-2.281	-1.997	-2.399	-2.257	-2.437	-2.459	-2.308	-2.327	-2.726
Rank		12	5	3	9	11	6	10	4	2	8	7	1
	p-value	0.040	0.008	0.008	0.012	0.024	0.009	0.013	0.008	0.008	0.011	0.010	0.003

Table 6 Out-of-sample forecast results for the period July 1, 2008 to December 30, 2008. Results are reported for examined models and the considered performance evaluation measures MSE, RMSE, MAE, MAPE, Theil U as well as results for Diebold-Mariano test with RW model as benchmark.

Given the confidence levels of 95% and 99% and a total of 127 days during the out-of-sample period, one would expect approximately 6.35, respectively, 1.27 VaR exceptions. To test for the appropriateness of the considered VaR models,

We apply a test that is based on the actual number of observed exceptions versus the expected number of exceptions, see e.g. Hull (2007). The test uses a binomial distribution such that given a true probability *p* of an exception, the probability of the VaR level being exceeded *m* or more days is:

$$\sum_{k=m}^{n} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{22}$$

A similar reasoning applies to the case where the number of VaR violations *m* is lower than the expected number of exceptions. The probability of m or less exceptions is:

$$\sum_{k=0}^{m} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{23}$$

Based on these quantities it is easy to derive *p*-values for a correct VaR model specification given the number of exceptions that were actually observed.

We find that the random walk model performs rather poorly both for the 95% and 99% VaR. For the long position, we observe 18, respectively 16 VaR exceptions corresponding to a failure rate of 14.2% and 12.6% that is substantially higher than the expected 5% and 1% under the assumption of a correct model specification. Similar results are obtained for holding a short position where the fraction of VaR exceptions is approximately 11% and 9.4%, respectively. Thus, as indicated by the *p*-values, for both 95% and 99% VaR levels, the model is significantly rejected.

While most of the models provide clearly less VaR violations than the RW model, only few of them are not rejected by the test for at least one of the two considered confidence levels. The HM and OLS model also significantly underestimate the risk, and yield too many exceptions for both long and short positions in particular at the 0.01 level. On the other hand, the three MA models yield a very small number of VaR violations, but the estimates are too conservative. As indicated in Table 7, for the long position, each MA model only yields one exception

Tables 7: Out-of-Sample Results for Periods 2

	RW	HM	MA(20)	MA(40)	MA(120)	OLS	AR(5)	MAD(5)	ARMA(1,1)	EWMA	GARCH(1,1)	GARCH-M	TARCH
Long Position α = 5% (expected violations 6.35)													
Violations	18	12	1	1	1	9	3	5	2	2	2	2	4
Failure rate	0.142	0.094	0.008	0.008	0.008	0.071	0.024	0.039	0.016	0.016	0.016	0.016	0.031
p-value	0.0002	0.0113	0.0114	0.0114	0.0114	**0.1046**	**0.1163**	**0.3860**	0.0442	0.0442	0.0442	0.0442	**0.2338**
Long position α = 1% (expected violations 1.27)													
Violations	16	8	0	0	1	7	2	4	1	0	1	1	2
Failure rate	0.126	0.063	0.000	0.000	0.008	0.055	0.016	0.031	0.008	0.000	0.009	0.008	0.016
p-value	0.0000	0.000	**0.2790**	**0.2790**	**0.6370**	0.0000	**0.1352**	0.0093	**0.6370**	**0.2790**	**0.6370**	**0.6370**	**0.1352**

Tables 8: Out-of-Sample Results for Period 3

	RW	HM	MA(20)	MA(40)	MA(120)	OLS	AR(5)	MAD(5)	ARMA(1,1)	EWMA	GARCH(1,1)	GARCH-M	TARCH
Short position α = 5% (expected violations 6.35)													
Failures	14	12	0	0	2	9	1	1	0	0	1	0	1
Failure rate	0.110	0.094	0.000	0.000	0.016	0.071	0.008	0.008	0.000	0.000	0.008	0.000	0.008
p-value	0.0018	0.0113	0.0015	0.0015	0.0044	**0.1046**	0.0114	0.0114	0.0044	0.0044	0.0114	0.0044	0.0114
Short position α = 1% (expected violations 1.35)													
Failures	12	9	0	0	1	6	1	0	0	0	0	0	1
Failure rate	0.094	0.071	0.000	0.000	0.008	0.047	0.008	0.000	0.000	0.000	0.000	0.000	0.008
p-value	0.0000	0.0000	**0.2790**	**0.2790**	**0.6370**	0.0003	**0.6370**	**0.2790**	**0.2790**	**0.2790**	**0.2790**	**0.2790**	**0.6370**

Table7 & 8: Number of VaR violations for the considered models assuming a short position in gold. Results are reported for 1-day 95% and 99%-VaR. The p-values refer to a test based on Hull (2007) where the null hypothesis is an appropriate specification of the VaR model. Bold letters indicate models that were not rejected at the 5% significance level.

at the 95% VaR level leading to a rejection of the models even at the 0.10 significance level. Almost the same results are obtained for holding a short position in the gold market where the 95%-VaR estimates are also too conservative, so all MA models are rejected. Note however, that the models are not rejected for the 99%-VaR level since only a very small number of exceptions are expected at this level. Similar results are obtained for the ARMA, EWMA and two models with conditional variance GARCH(1,1) and GARCH-M model.

These models only yield two exceptions at the 95% level and zero or one exception at the 99% level for a long position: for a short position, only the GARCH(1,1) model yields one exception at the 95% confidence level. The VaR estimates of these models are too conservative for the considered time period such that all models are rejected at the 5% significance level. The MAD(5) model gives too many exceptions at the 95% confidence level for a long position in gold, while it performs reasonably well at the 99% level for short positions.

The best results – at least for long positions - are obtained for the AR(5) model and again for the threshold conditional volatility TARCH model. These models seem to provide adequate one-day-ahead risk forecasts for long positions and cannot be rejected for any of the considered confidence levels. Considering short positions, the models seem to provide estimates that are overly conservative and yield only one exception at the 95% and no exception at the 99% confidence level. Still, given the reasonable performance of the AR(5) and GARCH models for long positions, they could be considered as being most appropriate in terms of providing VaR forecasts. Overall, we conclude that there was no clear winner with respect to providing one-day ahead Value-at-Risk forecasts.

5. Summary and Conclusions

In this paper we investigate the modelling of volatility dynamics of gold market returns in London. Gold markets are usually considered as a safe haven and investments into this class of assets have been very popular, in particular, since the global financial crisis. Therefore, appropriate models for volatility dynamics in these markets are of great interest to both investors and hedgers. While there are a number

of recent studies examining volatility and Value-at-Risk (VaR) measures in financial and commodity markets, none of them focuses in particular on the gold market. Compared to the numerous studies on volatility modelling and forecasting focused on equity and commodity markets in general, we provide a pioneering study on the volatility of this important market. We contribute to the literature by using a large number of statistical approaches in order to model and forecast the daily volatility and Value-at-Risk in the gold spot market. Hereby, we distinguish between different time horizons including a sub-period of continuously but only slightly increasing gold prices, a sub-period of substantially increasing gold prices and, finally, a sub-period of high volatility in the gold market. Both in-sample and out-of-sample forecasts are evaluated using appropriate forecast evaluation measures.

For in-sample forecasting, the class of TARCH models provided the best results among the tested models. Interestingly, the performance of a GARCH (1,1) model, that is generally supported by empirical studies for volatility modelling in financial markets (Akgiray, 1989; Franses and van Dijk, 1996), was only ranked in the middle of all models in our study. For out-of-sample forecasting, results were not that clear-cut and the order and specification of the models was found to be an important factor in determining the model's performance. VaR for traders with long and short positions were evaluated by comparing actual VaR exceptions to theoretical rates. For this task a simple AR as well as a TARCH model performed best for the out-of-sample period. We also find that most models were able to significantly outperform a benchmark random walk model both in the in-sample and the out-of-sample forecasting. However, none of the considered models performed significantly better than the rest with respect to all of the considered criteria.

The out-of-sample period from July to December 2008 that has been tested in this study was one of the most volatile periods in the history of financial markets. As a result, the behaviour of the daily returns might be significantly different to previous periods and, also, possibly future periods. Thus, models that perform well in the considered out-of-sample period may well underperform in future periods, particularly when market conditions change. Second, though the study attempts to comprehensively investigate the volatility in the gold market by the means of using various models, it still only covered a small number of models available in this area.

For example, for models with conditional volatility, only three of the most widely used GARCH models were considered, leaving out a huge number of other GARCH model extensions. The flaws of VaR as a measure of risk along with the effectiveness of alternative risk measures such as expected shortfall, have been pointed out in the literature by e.g. Artzner *et al*. (1999). We leave the investigation of these issues to future work.

Author information: Stefan Trück is a Professor of Finance in the Department of Applied Finance and Actuarial Studies and Co-Director of the Centre for Financial Risk at Macquarie University. Email: Stefan.trueck@mq.edu.au. Kevin Liang is a graduate student of Macquarie University. He has extensive professional experiences in the finance industry and is currently working as a credit risk analyst. Email: kzyliang@yahoo.com.au.

References

Akgiray, V. (1989). Conditional heteroskedasticity in time series of stock returns: evidence and forecasts. *The Journal of Business, 62*(1): 55-80.

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

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

Bollerslev, T., Chou, R. Y., & Kroner, K. F. (1992). ARCH modeling in finance: A review of the theory and empirical evidence. *Journal of Econometrics, 52*(1-2) : 5-59.

Brooks, C. (1998). Predicting stock index volatility: can market volume help? *Journal of Forecasting, 17*(1) : 59-80.

Brooks, C., & Persand, G. (2002). Models choice and value at risk performance. *Financial Analysts Journal, 58*(5) : 87-97.

Brooks, C., & Persand, G. (2003). Volatility forecasting for risk management. *Journal of Forecasting, 22*(1) : 1-22.

Canarella, G., & Pollard, S. K. (2008). Modelling the volatility of the London gold market fixing as an asymmetric power ARCH. *The Journal of Applied Finance, 14*(5) : 17-43.

Chou, R. Y. (1988). Volatility persistence and stock valuation: some empirical evidence using GARCH. *Journal of Applied Econometrics, 3*(4) : 279-294.

Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review, 39*(4) : 841-862.

Christoffersen, P. F., & Diebold, F. X. (2000). How relevant is volatility forecasting for financial risk management. *Review of Economics and Statistics, 82*(1) : 12-22.

Danielsson, J., & de Vries, C. G. (2000). Value-at-risk and extreme returns. *Annales d'Economie et de Statistique*(60) : 11.

Diebold, F. X. (1998). *Elements of Forecasting*. Cincinnati, Ohio: South Western College Publishing.

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

Dimson, E., & Marsh, P. (1990). Volatility forecasting without data-snooping. *Journal of Banking and Finance, 14*(2-3) : 399-421.

Ding, Z., Granger, C. W. J., & Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance, 1*(1) : 83-106.

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica, 50*(4) : 987-1007.

Engle, R. F., Lilien, D. M., & Robins, R. P. (1987). Estimating time varying risk premia in the term structure: the ARCH-M model. *Eonometrica, 55*(2) : 391-407.

Franses, P. H., & van Dijk, D. (1996). Forecasting stock market volatility using (non-linear) GARCH models. *Journal of Forecasting, 15*: 229-235.

French, K. R., Schwert, G. W., & Stambaugh, R. F. (1987). Expected stock returns and volatility. *Journal of Financial Economics, 19*(1) : 3-29.

Giot, P., & Laurent, S. (2003). Market risk in commodity markets: a VaR approach. *Energy Economics, 25*(5) : 435-457.

Giot, P., & Laurent, S. (2003). Value-at-risk for long and short trading positions. *Journal of Applied Econometrics, 18*(6) : 641-663.

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

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

Henriksson, R. D., & Merton, R. C. (1981). On market timing and investment performance. II. Statistical procedures for evaluating forecasting skills. *Journal of Business, 54*(4) : 513-533.

Hentschel, L. (1995). All in the family, nesting symmetric and asymmetric GARCH models. *Journal of Financial Economics, 39*(1) : 71-104.

Hull, J. C. (2007). *Risk management and financial institutions*: Pearson Prentice Hall.

Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoskedasticity and serial independence of regression residuals. *Economics Letters, 6*(3) : 255-259.

Jorion, P. (1996). Risk^2: Measuring the risk in value at risk. *Financial Analysts Journal, 52*(6) : 47-56.

Jun, J. H. (2009) Global financial crisis and gold market. *Working Paper*.

Kupiec, P. (1995). Techniques for veryfying the accuracy of risk management models. *Journal of Derivatives, 3*(2) : 73-84.

McMillan, D., Speight, A., & Apgwilym, O. (2000). Forecasting UK stock market volatility. *Applied Financial Economics, 10*(4) : 435-448.

Morales, L. (2008). *Volatility spillovers on precious metals markets: the effects of the Asian crisis*. Paper presented at the European Applied Business Research Conference (EABR), Salzburg, Austria.

J.P. Morgan., RiskMetrics Monitor, Fourth Quarter, 1995.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: a new approach. *Econometrica, 59*(2) : 347-370.

Sadorsky, P. (2006). Modeling and forecasting petroleum futures volatility. *Energy Economics, 28*(4) : 467-488.

Stock, J. H., & Watson, M. W. (1998). A comparison of linear and non-linear univariate models for forecasting macroeconomic time series. *NBER Working Paper, 6607*.

Tully, E., & Lucey, B. M. (2007). A power GARCH examination of the gold market. *Research in International Business and Finance, 21*(2) : 316-325.