Master’s Thesis Financial Economics

Erasmus School of Economics, Erasmus University Rotterdam

Evaluating Multi-Horizon Volatility-Forecasting Performances of GARCH-Type Models

Author: Han-Ywan Hu

ERNA no.: 449758

Supervisor: Dr Rogier Quaedvlieg

Second assessor: Dr Esad Smajlbegovic

Date: 7 February 2019

Acknowledgements

I would like to express my sincere gratitude to Dr Rogier Quaedvlieg, my thesis supervisor, for his eternal patience, guidance, useful comments, and flexibility to let me pursue two banking internships at ABN AMRO and ING during my study. Besides, I would like to thank the second assessor for his courage to read this tedious paper. Furthermore, I would also like to thank my family, my dearest friends, and everyone else for their encouragement and support during this prolonged but rewarding journey. This accomplishment would have not been possible without them.

万事开头难。

冰冻三尺，非一日之寒。

Rotterdam, 7 February 2019

Han-Ywan Hu

Evaluating Multi-Horizon Volatility-Forecasting Performances of GARCH-Type Models

Han-Ywan Hu

Erasmus School of Economics, Erasmus University Rotterdam

7 February 2019

Abstract

This paper examines the forecast performances of several GARCH models at different horizons for equity indices from developed markets and emerging markets. It tries to find out which GARCH models are best for short-term volatility forecasts and which ones perform best at longer horizons. Furthermore, we also study the underlying features that drive the forecast performance of GARCH models at multiple horizons. We use the ARCH, GARCH, GJR-GARCH, CGARCH, APARCH, and GARCHX models to forecast out-of-sample conditional variances of S&P 500, HSI, IPC, and KOSPI. The forecast performances are evaluated with MSE and QLIKE. Additionally, we test the significance of forecast losses with the Diebold-Mariano test and Model Confidence Set (MCS). The empirical findings of this paper are mixed, since there is not a single model that shows superior performance in each case. The GARCHX model, which incorporates the VIX, tends to generate superior forecasts at multiple horizons for the S&P 500, whereas in most cases CGARCH and APARCH are preferred under the remaining indices. Overall, we can conclude that conjoining a precise volatility proxy of an index to an ordinary GARCH model significantly improves the forecast performance, yet there are other factors as well that may influence the predictive ability of models.

Keywords: Forecasting, ARCH, GARCH, GJR-GARCH, CGARCH, APARCH, GARCHX, Realised Volatility, VIX

JEL: C22, C52, C53, C58

Table of contents

List of tables

List of figures

§1. Introduction

"We really can't forecast all that well, and yet we pretend that we can, but we really can't."

— Alan Greenspan

Volatility forecasting of financial assets such as bonds and equities is a common practice for scholars and practitioners in the field of corporate finance.[1] As mentioned in Poon and Granger (2005), volatility forecasting has a significant role, for instance, in risk management, investment analysis, and pricing of derivatives. Especially in option pricing theory, the volatility of an underlying asset is one of the key parameters in the notorious Black-Scholes-Merton formula, where one can determine the theoretical value of an option contract (Figlewski, 1997). Besides its essential role in option pricing, it also has a key role in risk management where volatility is an element of Value-at-Risk models (Hull, 2012). Undoubtedly, finance professionals are faced with different types of financial risks on a daily basis. Therefore, a forward-looking volatility measure of asset prices over a holding period can be considered as a good starting point for evaluating investment risk (Poon & Granger, 2003).

From an academic perspective, many studies have conducted research on forecasting models and their predictive abilities. In addition, studies have directly compared the forecasting performances of economic variables of these forecasting models. For example, Poon and Granger (2003) delineate the results of forecast performances of various volatility-forecasting models from 66 studies. However, a sceptical reader may question the validity of mathematical and statistical models for forecasting purposes. Firstly, can we predict future movements of economic variables such as volatility of a financial asset exactly? Secondly, which model is considered as the best model for volatility forecasting? The first question sounds rather naïve since economists often regard the future as uncertain, which of course complicates the matter of making a precise prediction of the future. According to the renowned essayist Nassim Nicholas Taleb, extreme events such as the 9/11 terrorist attacks in the U.S. are unknown unknowns.[2] These rare events defy any prediction made by statistical models because these events are one-off (Kay, 2007). Thus, future movements of economic variables are unlikely to be predicted exactly by existing forecast models. Then, the answer to the second question is rather ambiguous due to the mixed results of forecast performances reported in previous studies, which will be explained in more detail in the next section. Moreover, asset returns often do not contain sufficient

[1] In this paper, the words volatility, σ, and variance, σ^2, are interchangeable. This should not lead to conceptual confusion since these two statistical measures are related to each other through a monotonic transformation.

[2] Taleb labels unknown unknowns as "*black swans*".

information to pinpoint a generic forecast model as "best" (Hansen et al., 2003). Therefore, it is hard to find a clear winner when it comes to forecasting economic variables.

The global financial crisis of 2007-2008 clearly revealed that heavy reliance on complex mathematical and statistical models do not necessary lead to correct predictions of future market movements. Despite the frequent criticism of forecasting models, it is important to note that economic and financial outlooks are necessary for executives, institutional investors, investment bankers, and policy-makers because they need to anticipate certain shocks. They need mathematical and statistical tools to understand the nature of uncertainty to make well-informed decisions. Therefore, many mathematical models are still used for forecasting purposes nowadays.

The aim of this paper is mainly to evaluate the forecast performance of several GARCH-type models at different horizons, also known as multi-horizon forecasting. Further, it studies the underlying features that drive the forecasting performance of these GARCH models at multiple horizons. Based on this, the following research questions can be formulated: **(1)** *"Which of the following GARCH-type models generate the best volatility forecasts in the short term: ARCH, GARCH, GJR-GARCH, CGARCH, APARCH, or GARCHX?"*, **(2)** *"Which of the following GARCH-type models generate the best volatility forecasts in the medium term: ARCH, GARCH, GJR-GARCH, CGARCH, APARCH, or GARCHX?"*, and **(3)** *"What key features drive the forecast performance of these GARCH-type models at multiple horizons?"* The abovementioned research questions can be answered through evaluations of forecast results of these models at multiple horizons, and by using several goodness-of-fit measures to determine the model's significance and predictive ability. Further, this study uses equity indices from both developed markets and emerging markets.

Why is it worth to examine the abovementioned research questions? Firstly, from institutional investors' perspective, it is important to know which forecast models are best applicable to different types of financial assets or even within the same asset class. Secondly, it is important to examine and compare the predictive abilities of both simple and sophisticated models. This paper studies the volatility-forecasting performances and abilities of the ARCH, GARCH, GJR-GARCH, CGARCH, APARCH, and GARCHX models. Finally, this study adds to existing literature by comparing the volatility-forecasting models' performances of developed and emerging markets equity indices over horizons ranging from one day to one quarter. Furthermore, it examines the underlying features of these GARCH-type models that drive the forecast performance at different horizons.

§2. Literature

Previous studies related to multistep forecasting will be examined in this section, thus providing us with more insights on volatility forecasting at different horizons and the well-known empirical regularities of asset return volatility.

2.1. Multi-horizon forecasting

Throughout the years, numerous studies have done research on the use of distinct forecast models and their ability to generate multistep forecasts of economic variables. For example, Marcellino et al. (2006) compared out-of-sample forecasts of 170 different macroeconomic variables at multiple horizons, meanwhile Clark and McCracken (2005) focused their empirical study on a single macroeconomic variable, namely core CPI inflation. Other studies like Ghysels et al. (2009), Brownlees et al. (2011) and Quaedvlieg (2018) focused their research on volatility forecasting of securities by using various forecast models. Moreover, Brownlees et al. (2011) also investigated how different forecasting procedures, e.g. forecast window length, innovation distribution, and frequency of model parameter re-estimations, could affect the predictive ability of those GARCH models.

Surprisingly, consensus regarding which forecast model performs the best has not been reached yet. For instance, Hansen and Lunde (2005) analysed more than 300 GARCH-type models in terms of their ability to predict conditional variances of the Deutsche Mark-USD currency rate and IBM stock returns. According to the authors, specifying the standard GARCH(1,1) model into a more complex form does not necessarily lead to outperformance compared to the standard form. However, as reported in the extensive literature review by Poon and Granger (2003), some empirical studies have demonstrated evidence that even a simple historical volatility model, e.g. historical moving average or random walk model, can beat a complex GARCH-type model.

There are two common methods to conduct multistep forecasts. One could use either the direct or the iterated approach, which is also called recursive forecast. The former uses a horizon-specific forecast model in which the response variable is the h-steps ahead direct forecast at a certain time, whereas the latter uses one-step ahead forecasts of a model to determine the h-steps ahead forecast recursively (Marcellino et al., 2006; Ghysels et al., 2009). In other words, h-steps ahead iterated forecasts are based on a function of lagged one-step ahead forecasts. Interestingly, some previous studies have demonstrated that multistep iterated forecasts tend to outperform direct forecasts in terms of forecast accuracy, although improvements are relatively small for some horizons, see Marcellino et al. (2006), Proietti (2011), and McCracken and McGillicuddy (2017).

2.2. Empirical regularities of asset return volatility

Finance literature has documented several important empirical regularities regarding historical asset returns and market volatility. For example, one of the characteristics is that asset return volatility tends to persist and cluster (Poon & Granger, 2005). Another important characteristic is the asymmetric nature of asset return volatility. The next subsubsections will briefly introduce some of these empirical regularities from previous literature.

• ***2.2.1. Fat-tailed distributions***

As documented in Mandelbrot (1963), Fama (1965), and Bollerslev et al. (1994), asset returns tend to be distributed in a leptokurtic way, i.e. a return distribution with positive excess kurtosis that contains more observations in the extreme tails (Tsay, 2005). Similar results for stock index returns can be found in **Figure 1** (see **Section 4.1**) of this paper. This empirical regularity is also known as fat-tailed distribution, or heavy tails, in finance literature.

• ***2.2.2. Volatility clustering***

Other empirical regularities are clustering of volatility and persistence of asset returns (Poon & Granger, 2005). An example of volatility clustering is provided in **Figure 2**, where periods of relatively high volatility tend to cluster with each other, and low-volatility periods tend to be trailed by periods of relatively low volatility (Mandelbrot, 1963). In other words, there are cycles of substantial dispersion in asset returns and cycles of relative tranquillity. Although the figure illustrates hefty shocks in asset return volatility over time, volatility tends to be mean reverting in general (Engle & Patton, 2001). According to Engle (1993), if substantial adjustments in markets tend to be tracked by similar-sized changes, in either direction, then volatility must be predictably high after hefty changes.

• ***2.2.3. Long-memory property***

Persistence in volatility, or long memory, means that volatility shocks tend to decay slowly over time. In other words, volatility is said to be persistent if today's asset return not only has a significant impact on conditional variance forecast of the next period, but also on the variance estimates many periods in the future (Engle & Patton, 2001). Ding et al. (1993) and Ding and Granger (1996) have demonstrated that the long-memory property exists for some speculative asset returns, such as the S&P 500 and the DM/USD currency pair. Although speculative asset returns hold slight autocorrelation, the absolute returns and their power transformations tend to be serially correlated (Conrad et al., 2011). To detect this phenomenon, autocorrelations of variance estimates can be used to figure out the persistence level of volatility. An example of this is demonstrated in **Figure 3**.

• *2.2.4. Leverage effect*

The leverage effect refers to the tendency for asset price adjustments to be inversely correlated with asset volatility changes (Bollerslev et al., 1994; Andersen et al., 2006). In other words, declining asset prices are often accompanied by rising volatility, and vice versa (Aït-Sahalia et al., 2013). The impact of asset returns on volatility tend to be relatively weaker in bull markets than in bear markets, i.e. volatility is asymmetric over time (Poon & Granger, 2005). A common explanation for this effect is the change in a company's financial leverage, or its debt-to-equity ratio. In case when the enterprise value of a firm falls, the company with debt and equity outstanding typically becomes mechanically more levered (Bollerslev et al., 1994; Aït-Sahalia et al., 2013). Yet, the magnitude of the impact of a stock price drop on future volatilities seems too substantial to be explained merely by adjustments in company's financial leverage (Bollerslev et al., 2006; French et al., 1987). Still, according to Christie (1982), the inverted relationship between volatility and stock price is to a substantial degree attributable to financial leverage. Christie (1982), Bouchaud et al. (2001), and Aït-Sahalia et al. (2013) have tested the leverage effect empirically.

• *2.2.5. Volatility-spillover effect*

In the past decades, globalisation of financial markets and financial integration have led to intertwined capital markets on a massive scale. Nowadays, global markets are becoming increasingly interdependent, meaning that information flows instantaneously from one market to another, and thus affecting the securities market abroad. This phenomenon is also known as volatility spillover in finance literature. Previous studies have documented volatility-spillover effects in distinct financial markets. For instance, Hamao et al. (1990), Susmel and Engle (1994), Koutmos and Booth (1995), Ng (2000), and Baele (2005) studied the price volatility spillovers between various stock markets, meanwhile Skintzi and Refenes (2006) and Christiansen (2007) conducted their research mainly on the European bond markets. In addition, Tse (1999) and Yarovaya et al. (2017) investigated several stock index futures markets for spillover effects.

• *2.2.6. Commonality in volatility changes*

Another empirical regularity that is widely reported is the commonality in volatility changes, or the comovements in asset return volatility across distinct financial assets and even across international financial markets (Bollerslev et al., 1994). For example, changes in the volatility of equities and bonds tend to comove in the same direction (Schwert, 1989). This empirical regularity is closely linked with volatility-spillover effects.

§3. Methodology

3.1. Volatility proxies

A common issue in volatility forecasting is that conditional variance is not directly observable, which complicates the evaluation and comparison of predictions (Andersen & Bollerslev, 1998). Fortunately, there are ways to resolve this issue partially if an unbiased estimator of variance is available (Patton, 2011b). Empirical studies have proposed numerous proxies for conditional variance such as squared return, realised variance, and range-based variance, also called high-low range. Nevertheless, using a conditionally unbiased variance proxy does not necessarily result in the same outcome as if the true variance were used, as volatility proxies tend to be rather noisy in practice. Moreover, previous studies, such as Andersen et al. (1999) and Poon and Granger (2003), noted that a few outliers in sample data might substantially affect the outcomes of forecast evaluation and comparison tests. One way to tackle this issue, for example, is to use loss functions that are relatively less sensitive to outliers. However, Patton (2011b) found that many of those loss functions are not robust and thus can potentially result in erroneous rankings of conditional variance estimates. Throughout the years, no clear consensus has been reached yet on the treatment of outliers by academics (Poon & Granger, 2003).

This paper uses realised variance and range-based variance as volatility proxies. These estimators are described in the following subsubsections. The reason why we use these volatility proxies is because of forecast evaluation later. Before we move on to these subsubsections, we will first briefly discuss about the squared return.

- ***3.1.1. Squared return***

According to Patton (2011a), the daily squared return of a financial asset is considered as an unbiased volatility proxy that is simple and widely available for investors and researchers. However, despite the convenience of this proxy measure, squared return is also considered as a very noisy indicator of volatility (Andersen et al., 2001). While squared return is an unbiased estimator for the latent variable of interest, the idiosyncratic error term causes the inaccuracy of this proxy (Andersen & Bollerslev, 1998). As demonstrated in Lopez (2001), the squared error, ε_t^2, is 50% greater or smaller than the conditional variance, σ_t^2, for nearly three-fourths of the time. As a result, the squared return approach often leads to severe volatility forecasts. Hence, this method will not be used in this paper as it is expected to underperform other variance proxies.

• *3.1.2. Realised variance*

An alternative proxy for volatility is realised variance, or realised volatility (henceforth RV), which is based on intraday returns. The availability of high-frequency data led to better estimates of daily volatility, due to increased number of intraday observations. As described in Christoffersen (2012) and Patton (2011b), the realised variance is simply the sum of all the intraday squared log returns. Mathematically, it can be written as

$$RV_t^m = \sum_{i=1}^{m} r_{t-1+i/m}^2, \tag{1}$$

where m is the amount of observations available within a day assuming that those observations are equally spaced, and i denotes the ith intraday return on day t. Contrary to the daily squared return, RV is a proxy that can illustrate the daily variance in multiple ways through changes in the time interval, i.e. one can determine realised variance by using intraday observations that are sampled at, for instance, k-minute or hourly intervals. As a result, numerous estimates of RV are available. According to Poon and Granger (2005), from all the available RV estimates, studies have demonstrated that 5-minute or 15-minute interval produces the best results. This paper uses 5-minute and 10-minute realised volatilities since these time series are retrievable from Oxford-Man Institute's realised library.

Like the daily squared return, RV is a valid proxy of the daily conditional variance (Patton, 2011b). What characterises this volatility estimator is that in absence of market microstructure noise it is a more efficient estimator than the daily squared return (Martens & van Dijk, 2007). Moreover, RV is extremely persistent, which implies that volatility may be forecastable at short horizons if the information in intraday returns is used (Christoffersen, 2012). According to Andersen et al. (2006), the key feature of RV is that it provides a consistent nonparametric estimation of the variability of price that has occurred over a given discrete interval.

Unfortunately, the presence of market microstructure frictions significantly affects the estimation of realised variance. For example, in a univariate case, a bounce in the bid-ask spread of a financial asset tends to increase the measured volatility of the asset's high-frequency intraday returns. As a result, the volatility proxy under **Eq. (1)** tends to be upwardly biased due to the cumulated bias in intraday returns (Alizadeh et al., 2002). Market microstructure noise causes serial correlation in the intraday returns and therefore makes the RV estimator biased (Hansen & Lunde, 2006). To diminish the impact of market microstructure noise on volatility estimate, RV estimator can be constructed at a moderate frequency. However, according to Christensen and Podolskij (2007), this leads to loss of information.

• 3.1.3. Range-based variance

In presence of market microstructure noise, the construction of periodic realised volatilities becomes an arduous task, for example, due to inconsistent or lack of clean high-frequency data. In contrast to *RV* estimator, the range-based variance proxy, or range proxy (henceforth *RP*), is relatively less susceptible to market microstructure frictions (Christoffersen, 2012). This estimator reveals relatively more information about the actual volatility than intraday returns sampled at fixed intervals, since the daily range values are formed from the entire price process (Christensen & Podolskij, 2007). Similar to Christoffersen (2012), assuming that log returns are normally distributed with mean and variance equal to zero and σ^2, respectively, the range proxy can be defined as

$$RP_t = \frac{1}{4\ln(2)}\left(\ln\left(\frac{P_t^{High}}{P_t^{Low}}\right)\right)^2, \qquad (\mathbf{2})$$

where P_t^{High} and P_t^{Low} are the highest and lowest observed prices at time t, respectively. The ratio P_t^{High}/P_t^{Low} illustrates the high-low range of an asset. In addition, the constant $1/(4\ln(2))$ is a scaling factor, which represents the second moment of range of a Wiener process (Bannouh et al., 2009).

However, **Eq. (2)** does not incorporate the financial asset's opening and closing prices on day t, despite that they may improve the estimation of 24-hour volatility. Therefore, a more accurate range proxy is available, assuming again that log returns are normally distributed, namely

$$RP_t^* = \frac{1}{2}\left(\ln\left(\frac{P_t^{High}}{P_t^{Low}}\right)\right)^2 - (2\ln(2)-1)\left(\ln\left(\frac{P_t^{Close}}{P_t^{Open}}\right)\right)^2. \qquad (\mathbf{3})$$

Note that P_t^{Close} and P_t^{Open} are the observed closing and opening prices at time t, respectively. Although the inclusion of opening and closing prices can improve the variance estimator, the gains are not necessarily realised in practice because of the effects of market microstructure noise on asset prices (Alizadeh et al., 2002; Martens & van Dijk, 2007). Yet, we will use these two *RP* measures in this research as alternatives to the realised variance estimators.

As stated in Alizadeh et al. (2002) and Bannouh et al. (2009), the range-based approach has been known as a volatility estimator for a long time. Parkinson (1980) has demonstrated that the extreme value method, or high-low range, is noticeably more efficient as a variance estimator compared to the squared return method. Other studies have exploited this result for development of volatility proxy based on intraday high-low prices that appears to be far more efficient than the realised variance, see Martens and van Dijk (2007) and Christensen and Podolskij (2007).

3.2. Forecasting models

Throughout the years, literature has propounded numerous forecast models that are to a certain extent applicable to applied economic and financial research. As discussed in Christoffersen (2012), the goal of a volatility-forecasting model is to predict future variance as precisely as possible to apply necessary risk measures. In the past decades, many researchers have evaluated and analysed the forecasting performance of distinct volatility-forecasting models, ranging from simple moving average to sophisticated stochastic models. An extensive review of 93 empirical studies can be found in Poon and Granger (2003). We will focus on six GARCH-type models, namely the ARCH, GARCH, GJR-GARCH, CGARCH, APARCH, and GARCHX models. We will discuss these forecast models in more detail in the next subsubsections.

• ***3.2.1. ARCH***

In 1982, Engle introduced a new class of stochastic process called the Autoregressive Conditional Heteroskedasticity (ARCH) process, which allows the conditional variance to be time varying as a linear function of lagged errors, leaving the unconditional variance constant over time (Engle, 1982; Bollerslev, 1986). ARCH was one of the first econometric models that provided a convenient way to model conditional heteroskedasticity in variance. First, to model an ARCH process, let ξ_t denote the disturbance term, which depends on a stochastic component z_t and a time-varying standard deviation σ_t (Nelson, 1991). Mathematically, it can be written as

$$\xi_t = \sigma_t z_t,$$

where $z_t \sim$ i.i.d. $\mathcal{N}(0,1)$. By definition, ξ_t is serially uncorrelated with mean zero and conditional variance equal to σ_t^2. In mathematical terms, $\xi_t|\psi_{t-1} \sim \mathcal{N}(0, \sigma_t^2)$ where ψ_t denotes the set of all information available at time t (Engle, 1982). The conditional variance, σ_t^2, is modelled as follows

$$\sigma_t^2 = \omega_0 + \sum_{j=1}^{q} \alpha_j \xi_{t-j}^2, \tag{4}$$

where $\omega_0 > 0$ and $\alpha_j \geq 0 \; \forall j \in \{1,2, \dots, q\}$. In this model, σ_t^2 is expressed as a linear function of lagged squared errors where recent disturbances matter more for current period's conditional variance than distant errors (Bollerslev et al., 1992). Hence, this model is referred as the $ARCH(q)$ model.

The main strengths of the ARCH model are its simplicity and ability to incorporate well-known empirical regularities such as heavy tails, volatility clustering, and mean reversion (Bera & Higgins, 1993). **Eq. (4)** captures the conditional and unconditional variances in σ_t^2 and ω_0, respectively. The latter ensures that the model incorporates the mean reverting characteristic of the data. Despite its novelty at that time, the model has several limitations as well. First, ARCH requires a high order of lag variables to capture the dynamic behaviour of conditional variance in practice, see Bollerslev (1986), Bollerslev et al. (1992), Bera and Higgins (1993), and Bollerslev et al. (1994). This is simply a computational burden for the modeller due to the estimations of many parameters that are subject to inequality restrictions. Second, Engle's model assumes the conditional variance to be a linear function of lagged squared error terms that does not take into account the leverage effect (Nelson, 1991; Bera & Higgins, 1993). In other words, σ_t^2 is affected only by the magnitude and not by the sign of ξ_{t-j}. Accordingly, the ARCH model is not able to incorporate the asymmetric behaviour of volatility shocks.

• ***3.2.2. GARCH***

Not long after the introduction of Engle's ARCH model, Bollerslev proposed a more generalised version of this model in 1986, also called the Generalised Autoregressive Conditional Heteroskedasticity (GARCH) model. Bollerslev's GARCH model is an extension of Engle's ARCH model that allows for more flexible lag structure, i.e. the GARCH model requires less parameters to be estimated than the ARCH model to forecast conditional variances (Bollerslev, 1986). Since the properties of the ARCH and GARCH models are reasonably similar, $GARCH(p, q)$ is defined as

$$\sigma_t^2 = \omega_0 + \sum_{i=1}^{p} \varphi_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \alpha_j \xi_{t-j}^2, \tag{5}$$

where ω_0, φ_i, and α_j are nonnegative parameters. For a stable conditional variance process, we require the weights to be $\sum_{i=1}^{p}\sum_{j=1}^{q}(\varphi_i + \alpha_j) < 1$ (Hull, 2012). In this case, the conditional variance, σ_t^2, is a linear function of past estimates of variance rates and past squared errors, where more recent observations of both parameters get relatively higher weights compared to more distant observations. Interestingly, the conditional variance of ξ_t behaves like an ARMA process (Enders, 2015). Notice that **Eq. (4)** is a special case of **Eq. (5)**.

As noted by Nelson (1991), GARCH models elegantly capture the volatility clustering in asset returns. Another important feature is that these models can be fitted to data that have excess kurtosis (Franses & Ghijsels, 1999). Furthermore, these models recognise the mean-reverting dynamics of asset return volatility, which implies that the variance tends to converge to a long-run average variance rate (Hull, 2012). More importantly, the GARCH model is often considered as a parsimonious model since it uses less lag variables compared to the ARCH models in empirical applications (Enders, 2015). The above features accentuate the key strengths of this econometric model.

Unfortunately, the GARCH model has its limitations as well. Similar to the ARCH model, the standard GARCH model lacks the ability to incorporate the volatility asymmetry of positive and negative returns (Nelson, 1991). Another limitation results from the nonnegativity constraints on the variance equation parameters, see **Eq. (5)**, which rule out any stochastic oscillatory behaviour in the σ_t^2 process and can complicate the estimation of GARCH models. Lastly, the interpretation of the persistence of volatility shocks may be questionable. If volatility shocks persist continuously, they may stir the entire term structure of risk premia, and therefore are likely to affect significantly the investment in tangible assets (Poterba & Summers, 1986).

• *3.2.3. GJR-GARCH*

As stated previously, the standard GARCH model does not incorporate the asymmetric dynamics of asset return shocks. Fortunately, a commonly used asymmetric GARCH-type model that captures this leverage effect phenomenon is called the GJR-GARCH model, proposed by Glosten et al. (1993). Compared to the normal GARCH model, GJR-GARCH is a richer model since it augments the GARCH model with a dummy variable I_{t-j} that takes either a value of one or zero (Brailsford & Faff, 1996). Hence, I_{t-j} is mathematically defined as

$$I_{t-j} = \begin{cases} 1 & \Leftarrow \xi_{t-j} < 0 \\ 0 & \Leftarrow \xi_{t-j} > 0. \end{cases}$$

As a result, we can specify the $GJR\text{-}GARCH(p, q)$ model as

$$\sigma_t^2 = \omega_0 + \sum_{i=1}^{p} \varphi_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \left(\alpha_j + \lambda_j I_{t-j}\right) \xi_{t-j}^2, \tag{6}$$

where ω_0, φ_i, α_j, and λ_j are nonnegative parameters. Besides, for a stable conditional variance process, we set $\sum_{i=1}^{p} \sum_{j=1}^{q} (\varphi_i + \alpha_j) < 1$. When the indicator variable $I_{t-j} > 0$, it simply implies that past negative return shocks have a larger impact on the future conditional standard deviations than past positive return shocks (Andersen et al., 2006). Note that **Eq. (6)** turns into **Eq. (5)** when $I_{t-j} = 0$.

• *3.2.4. CGARCH*

Another GARCH-type model that we will introduce is the component GARCH (henceforth CGARCH) model. This model tries to decompose volatility into a short-run component and long-run component (Engle & Lee, 1999; McMillan & Speight, 2004). Shocks of the long-run volatility component are highly persistent, whereas shocks of the transitory component are relatively less persistent (Watanabe & Harada, 2006). Normally, in a standard GARCH model, the conditional variance tends to mean revert to its long-run variance component, ω_0, over time. In contrast, CGARCH allows mean reversion to a time-varying trend (Engle & Lee, 1999). Thus, the $CGARCH(p,q)$ model is specified as

$$\sigma_t^2 = \widetilde{\omega}_t + \sum_{i=1}^{p} \varphi_i\left(\sigma_{t-i}^2 - \widetilde{\omega}_{t-i}\right) + \sum_{j=1}^{q} \alpha_j\left(\xi_{t-j}^2 - \widetilde{\omega}_{t-j}\right), \tag{7}$$

where $\widetilde{\omega}_t$ and $\left(\sigma_{t-i}^2 - \widetilde{\omega}_{t-i}\right)$ are the time-varying trend and short-run component of σ_t^2, respectively. This time-varying long-run variance rate is formally defined as

$$\widetilde{\omega}_t = h_0 + x\widetilde{\omega}_{t-1} + y(\xi_{t-1}^2 - \sigma_{t-1}^2).$$

In this case, the forecasting error $(\xi_{t-1}^2 - \sigma_{t-1}^2)$ drives the time-dependent movement of the trend variable. Importantly, to achieve stationarity, we set $(\varphi + \alpha)(1 - x) + x < 1$, which in turn requires $x < 1$ and $(\varphi + \alpha) < 1$. As a result, the short-run component then converges to zero with powers of $\varphi + \alpha$, while the long-run component converges to $\widetilde{\omega}_t$ with powers of x (McMillan & Speight, 2004).

• ***3.2.5. APARCH***

Ding et al. (1993) proposed a general class of forecast model that includes Engle's ARCH, Bollerslev's GARCH, and five other GARCH-type models as special cases. This polyvalent forecast model is also called the Asymmetric Power Autoregressive Conditional Heteroskedasticity (APARCH) model. Thus, the standard $APARCH(p,q)$ model is formally defined as

$$\sigma_t^\delta = \omega_0 + \sum_{i=1}^{p} \varphi_i \sigma_{t-i}^\delta + \sum_{j=1}^{q} \alpha_j\left(\left|\xi_{t-j}\right| - \lambda_j \xi_{t-j}\right)^\delta \tag{8}$$

where ω_0, δ, φ_i, and α_j are nonnegative parameters, meanwhile parameter $\lambda_j \in (-1,1)$. Notice that δ is the Box-Cox transformation, or exponentiation, of the conditional variance and the asymmetric absolute residuals (Ding et al., 1993; Laurent, 2004). This transformation is convenient for linearising nonlinear models. Furthermore, parameter λ_j reflects the so-called leverage effect (Brooks et al., 2000). When $\lambda_j > 0$, past negative shocks have a deeper effect on σ_t than past positive shocks. Meanwhile, $\lambda_j < 0$ implies that past positive shocks have a greater impact on σ_t than past negative shocks.

The key strength of the APARCH model is its ability to incorporate well-known empirical regularities, such as volatility clustering, leverage effect, and long-memory property, due to the flexible structure of the model which nests at least seven different ARCH-type models (Laurent, 2004). By changing the underlying parameters of APARCH, we obtain the ARCH, GARCH, TS-GARCH, GJR-GARCH, TARCH, NARCH, and Log-ARCH models (Ding et al., 1993). However, despite the model's novelty and versatility, the APARCH model does not allow fractional integration of conditional variance. Consequently, conditional variance shocks either dispel exponentially or persist indefinitely (Degiannakis, 2004). Baillie et al. (1996) and Tse (1998) proposed extensions to the APARCH model to address this limitation by expressing the conditional variance equation as a function of lagged shocks that decay at a gradual hyperbolic rate.[3]

• *3.2.6. GARCHX*

Throughout the years, many extensions to the classic GARCH model, ranging from nonlinear asymmetric power modifications to multivariate class models, have been proposed in the finance literature. Just like the Capital Asset Pricing Model (Sharpe, 1964; Lintner, 1965; Mossin, 1966), Arbitrage Pricing Theory (Ross, 1976), and Black-Scholes-Merton Model (Black & Scholes, 1973; Merton, 1973), the GARCH model can be easily extended to a more complex form by adding external regressors into the equation. In this paper, we will extend the $GARCH(p,q)$ by adding a covariate linked to the Cboe Volatility Index, or VIX, and call it the $GARCHX(p,q,r)$ model.[4] The VIX Index reflects investors' expectation of the volatility of the S&P 500 over the next month, which is implied by the current prices of S&P 500 index options (Kambouroudis & McMillan, 2016).[5] Thus, it can be seen as a forward-looking volatility measure. Therefore, including VIX as an additional independent variable may significantly improve the forecastability of a standard GARCH-type model (Christoffersen, 2012). Earlier empirical studies by Fleming et al. (1995), Hol (2003), Blair et al. (2010), and Kambouroudis and McMillan (2016) provided supportive evidence for VIX in providing accurate volatility forecasts.

Our $GARCHX(p,q,r)$ model is formally defined as

$$\sigma_t^2 = \omega_0 + \sum_{i=1}^{p} \varphi_i \sigma_{t-i}^2 + \sum_{j=1}^{q} \alpha_j \xi_{t-j}^2 + \sum_{k=1}^{r} \upsilon_k \zeta_{t-k}, \tag{9}$$

where variable $\zeta_t = \frac{1}{252}\left(\frac{VIX_t}{100}\right)^2$ is defined as a variance proxy for VIX_t. The daily VIX variance proxy is in accordance with Hao and Zhang's (2013) approach. Further, υ_k is a weight of covariate ζ_t. Hence, the conditional variance equation of GARCHX is now a linear function of lagged variances, squared errors, and VIX variance proxy. Fortunately, we can estimate covariate ζ_{t-1} with univariate maximum likelihood estimation.

[3] Baillie et al. (1996) and Tse (1998) introduced the FIGARCH and FIAPARCH, respectively, which are fractionally integrated extensions of the GARCH-type models.
[4] Cboe: Chicago Board Options Exchange.
[5] The Cboe VIX is often called "investor fear gauge" by financial news media. According to Cboe's website, the VIX Index can be considered as the world's premier barometer of investor sentiment and market volatility.

3.3. Model fitting & forecast evaluation

To assess the quality and predictive ability of forecast models, one should consider several tests for both short- and medium-term predictions of conditional variance. Therefore, we will discuss more about goodness-of-fit and forecast evaluation tests in the following subsubsections.

• *3.3.1. Information criteria tests*

A typical goodness-of-fit test indicates whether a forecast model is statistically significant, but it does not tell us how we should select between these models (Bozdogan, 1987; Kuha, 2004). Neither does it penalise for potential overfitting when additional parameters are added to the model. Fortunately, there are better equipped tests that avoid these types of issues such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), see Burnham & Anderson (2004). These in-sample fitting tests penalise for the number of model parameters to avoid possible overfitting issues. AIC and BIC are used as relative quality measures for selecting the best-fitting model amongst others. The forecast model with the smallest information criterion coefficient is preferred since it fits best with the in-sample data. Thus, the AIC is formally defined as

$$AIC = 2k - 2\ln(\hat{\theta}), \qquad (\mathbf{10})$$

where k is the number of model parameters and $\hat{\theta}$ is the maximum value of the forecast model's likelihood function (Bozdogan, 1987), meanwhile the BIC is defined as

$$BIC = \ln(n)\, k - 2\ln(\hat{\theta}), \qquad (\mathbf{11})$$

where n describes the quantity of sample information (Kuha, 2004).

• *3.3.2. Mean squared error*

Similar to Brownlees et al. (2011), Bollerslev et al. (2016), and Quaedvlieg (2018), we will use the mean squared error (henceforth MSE), or quadratic loss, measure to assess our conditional variance forecasts. As discussed in Patton (2011b), mean squared error is a robust loss function. According to Lopez (2001), MSE is a commonly used forecast evaluation tool for both in-sample and out-of-sample forecasts. Simply put, MSE is the average squared difference between the realised conditional variance and the corresponding volatility forecast. Mathematically,

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{VAR}_{t+i} - \hat{\sigma}^2_{t+h|t}\right)^2, \qquad (\mathbf{12})$$

where $\widehat{VAR}_{t+i}$ and $\hat{\sigma}^2_{t+h|t}$ are the unbiased ex-post estimate of conditional variance and h-step(s) ahead forecast of conditional variance, respectively. Squaring the difference results in larger weights for possible outliers. Dismally, this makes MSE sensitive to potential outliers. Eventually, the forecast model with the lowest MSE loss coefficient is preferred. In our case, we scale the loss coefficients for convenience sake by taking the square root of MSE, also known as root mean squared error (RMSE).

Unfortunately, basing forecast evaluation on the MSE approach results in two shortcomings. Firstly, despite that $\widehat{VAR}_{t+i}$ is an unbiased variance estimator, it is technically still an imprecise estimator (Lopez, 2001). Secondly, MSE penalises conditional variance estimates that are distinct from the realised conditional variance in a fully symmetrical way. It does not penalise for negative or zero estimates of the conditional variance (Bollerslev et al., 1994). In practice, however, one may reckon contrastingly about negative and positive forecast errors of the same magnitude, as underestimation of volatility may be costlier than overestimation by the same amount (Christoffersen, 2012).

• *3.3.3. Quasi-likelihood*

To allow for asymmetric loss when evaluating volatility forecasts, one can use the quasi-likelihood (henceforth QLIKE) loss function, which is another commonly used robust loss function in empirical studies. Like mean squared error, forecast model with the lowest QLIKE loss is preferred. The QLIKE is mathematically defined as

$$QLIKE = \frac{\widehat{VAR}_{t+i}}{\hat{\sigma}^2_{t+h|t}} - \log\left(\frac{\widehat{VAR}_{t+i}}{\hat{\sigma}^2_{t+h|t}}\right) - 1. \quad \textbf{(13)}$$

Note that this loss function depends on a relative forecast error, $\widehat{VAR}_{t+i}/\hat{\sigma}^2_{t+h|t}$, whereas MSE depends on an absolute forecast error (Brownlees et al., 2011). QLIKE will always penalise biased forecasts of conditional variance, i.e. underestimation of volatility, more heavily than MSE (Christoffersen, 2012). Here we first determine the model's QLIKE losses for the entire out-of-sample period and then estimate the model's mean loss.

According to Brownlees et al. (2011), there are reasons why someone prefers QLIKE to MSE for forecast comparison. First, since the QLIKE loss function depends on a relative volatility forecast error, the loss series is i.i.d. under the null, assuming that the forecast model is specified properly. MSE on the other hand, which depends entirely on $\widehat{VAR}_{t+i} - \hat{\sigma}^2_{t+h|t}$, will have a variance that is proportional to the square of the variance of asset returns (Patton, 2011b), and thus consists of high levels of serial dependence even under the null. Second, MSE has a bias that is proportional to the square of the true variance, whereas QLIKE bias is independent of the volatility level (Brownlees et al., 2011). Amidst severe market turmoil, sizable MSE loss coefficients will be a repercussion of high volatility regime without necessarily corresponding to the decay of a model's predictive ability. QLIKE avoids this vagueness, making it simpler to study forecast errors across different volatility regimes.

Nevertheless, both MSE and QLIKE loss functions do not tell anything about the statistical significance of forecast performances, making predictive comparisons incomplete (Diebold, 2015). The difference between loss coefficients may not be significantly different from zero (Choudhry & Wu, 2008). Therefore, to draw any meaningful inferences about the predictive ability of forecast models, we also need suitable significance tests.

• *3.3.4. Diebold-Mariano test*

The Diebold-Mariano (DM) test is a commonly used significance test for pairwise comparisons of competing forecasts at different horizons (Diebold & Mariano, 1995). Compared to the abovementioned loss functions, the DM test allows one to assess the statistical significance of apparent predictive superiority (Diebold, 2015). This method relies on assumptions made directly on the difference between two loss functions, also called loss differential. The DM test requires the loss differential to be covariance stationary, but it may not be strictly necessary in some cases (Diebold, 2015). For any loss function, the loss differential, $d_{ij,t}$, can be defined as

$$d_{ij,t} = L_{i,t} - L_{j,t}, \qquad (\mathbf{14})$$

with $L_{i,t}$ and $L_{j,t}$ as loss functions of forecast models i and j, respectively, at time t. In case when the disparity in the accuracy of two competing models is zero, then $\mathbb{E}(L_{i,t}) = \mathbb{E}(L_{j,t})$, or $\mathbb{E}(d_{ij,t}) = 0$. This corresponds with the equal predictive accuracy null hypothesis, which states that the population mean of a loss differential series is equal to zero (Diebold & Mariano, 1995). To test this null hypothesis, a simple asymptotic z-test can be used. Therefore, the DM test statistic can be obtained through

$$DM = \frac{\bar{d}_{ij}}{\hat{\sigma}_{\bar{d}_{ij}}}, \qquad (\mathbf{15})$$

where $\bar{d}_{ij} = \frac{1}{T}\sum_{t=1}^{T} d_{ij,t}$, or sample mean loss differential, and $\hat{\sigma}_{\bar{d}_{ij}}$ is a consistent estimator of the standard deviation of $\bar{d}_{ij}$. Under the null hypothesis, the DM test statistic has an asymptotic standard normal distribution (Harvey et al., 1997). Importantly, the DM statistic can be obtained by regression of the loss differential on a constant, using Newey-West standard errors (Diebold, 2015).[6]

However, just like other statistical metrics, the DM test is subject to flaws as well. Firstly, the initial DM test was found to be oversized for moderate numbers of sample observations (Harvey et al., 1997). Yet, the DM test is found considerably more versatile than other tests of equal forecast accuracy. When forecast errors are not normally distributed, the Morgan-Granger-Newbold test (1977)

---

[6] Newey-West standard errors are also called heteroskedasticity and autocorrelation consistent (HAC) robust standard errors (Newey & West, 1994).

and Meese-Rogoff test (1988) for equal MSE are severely missized in both large and small samples. The DM test, on the other hand, maintains approximately correct size for all but moderate-sized samples (Mariano, 2002). Secondly, reporting results of pairwise comparisons becomes rather laborious when the set of models increases, since one must perform $n(n-1)/2$ tests (Hansen et al., 2003).

• ***3.3.5. Model Confidence Set***

Contrary to the widely used Diebold-Mariano test, the Model Confidence Set (MCS), introduced by Hansen et al. (2003), allows for comparison of multiple forecast models at once. This model selection method is an innovative way to deal with the issue of selecting the "best" forecast model(s) using out-of-sample evaluation under a specified loss function. As defined in Hansen et al. (2011), an MCS, or $\mathcal{M}^*$, is a subset of a collection of candidate models, $\mathcal{M}^0$, which consists of superior forecast models for a given significance level. The set of superior forecast models is formally defined as

$$\mathcal{M}^* = \{i \in \mathcal{M}^0 : \mu_{ij} \leq 0, \forall\, j \in \mathcal{M}^0\}, \quad \textbf{(16)}$$

where $\mu_{ij} = \mathbb{E}(d_{ij,t})$ is finite and does not depend on t for all $i, j \in \mathcal{M}^0$. Note that $d_{ij,t}$ is the loss differential. The aim of this method is to determine the set of superior models, which can be done via a sequence of significance tests where models that are found to be significantly inferior to other models of $\mathcal{M}^0$ are eliminated (Hansen et al., 2011). Hence, MCS can be viewed as a sequential DM test (Quaedvlieg, 2018), or as a confidence interval of a parameter (Samuels & Sekkel, 2013). It is appealing to use a set of forecast models rather than an individual model, since there is no generic model that will consistently outperform other models in all conceivable scenarios.

This procedure is constructed from an equivalence test, $\delta_{\mathcal{M}}$, and an elimination rule, $e_{\mathcal{M}}$, that are assumed to have certain properties (Hansen et al., 2011). The former is used to test whether the null $H_{0,\mathcal{M}}: \mu_{ij} = 0, \forall\, i, j \in \mathcal{M}$ with $\mathcal{M} \subseteq \mathcal{M}^0$, while the latter identifies an inferior model from $\mathcal{M}$ and removes it when H_0 is rejected. Eventually, the surviving models end up in $\widehat{\mathcal{M}}^*$ after sequentially trimming $\mathcal{M}^0$ whilst holding the significance level, α, fixed at each procedural step (Hansen et al., 2003). The algorithm for constructing $\widehat{\mathcal{M}}^*_{1-\alpha}$ is as follows

Step I. Initiate the procedure by setting $\mathcal{M} = \mathcal{M}^0$.

Step II. Test null hypothesis by using $\delta_{\mathcal{M}}$ at level α.

- *a.* If $H_{0,\mathcal{M}}$ is not rejected, define $\widehat{\mathcal{M}}^*_{1-\alpha} = \mathcal{M}$.
- *b.* If $H_{a,\mathcal{M}}: \mu_{ij} \neq 0$ is accepted, then use $e_{\mathcal{M}}$ to remove forecast model i from $\mathcal{M}$, and reiterate *Step II* of this procedure.

To test the null hypothesis, we need to define $\delta_{\mathcal{M}}$, which is simply based on the well-known Diebold-Mariano test, see **Eq. (15)**. In this case, $\hat{\sigma}^2_{\bar{d}_{ij}}$ is a bootstrapped estimate of the variance of the sample mean loss differential that can be computed through a block bootstrap method of B resamples with a certain block length of $\mathcal{L}$ (Bernardi & Catania, 2018). Moreover, $e_{\mathcal{M}}$ is defined as

$$e_{max,\mathcal{M}} = \arg\max_{i \in \mathcal{M}} \frac{\bar{d}_i}{\hat{\sigma}_{\bar{d}_i}}. \tag{17}$$

Interestingly, this procedure yields p-values for all forecast models under consideration (Hansen et al., 2003). These MCS p-values are useful since then it is clear which forecast models are included in $\widehat{\mathcal{M}}^*_{1-\alpha}$ for any significance level. The p-value for $e_{\mathcal{M}_j} \in \mathcal{M}^0$, or forecast model j, is formally defined by $\hat{p}_{e_{\mathcal{M}_j}} = \max_{i \leq j} p_{H_0,\mathcal{M}_i}$, where $p_{H_0,\mathcal{M}_i}$ stands for the p-value that is associated with $H_{0,\mathcal{M}_i}$. Forecast models with relatively modest p-values are not probable to be contained in $\widehat{\mathcal{M}}^*_{1-\alpha}$, because $\hat{p}_i \geq \alpha$ for any $i \in \mathcal{M}^0$ in order to be part of $\widehat{\mathcal{M}}^*_{1-\alpha}$, see Hansen et al. (2011).

One of the benefits of the Model Confidence Set is that it acknowledges the limitations of the data (Hansen et al., 2003, 2011). In other words, informative data will lead to $\mathcal{M}^*$ that holds only the best model, whereas less informative data make it hard to differentiate between forecast models. Hence, this method is dissimilar to other model selection criteria that select a single model and disregard the surprisal of the underlying data. Another benefit is that this procedure makes it feasible to make statements regarding the significance that are valid in the conventional sense, a characteristic which is not satisfied by the frequently used approach of listing p-values from various pairwise comparison tests (Hansen et al., 2011). Furthermore, the MCS approach allows for the possibility that $\mathcal{M}^*$ contains more than one superior model. Lastly, $\mathcal{M}^*$ does not discard a forecast model unless it is found to be significantly inferior in comparison with other models (Hansen et al., 2003).

3.4. Forecasting procedure

Before we start using these GARCH-type models, the forecasting procedure needs to be outlined first. We split the data sample into two separate periods, namely in-sample and out-of-sample periods. The in-sample period starts from January 5, 2000, to August 6, 2004, whereas the out-of-sample starts from August 9, 2004, to March 27, 2018. The former results in 1,000 daily observations that covers periods of relatively large and small movements in volatility, meanwhile the latter has 3,078 out-of-sample observations. Hence, there are 4,078 daily observations in total (see **Table 1**). Model predictions that are based entirely on in-sample data are susceptible to potential overfitting, as in-sample forecast evaluations tend to spuriously indicate the existence of predictability when there is none (Inoue & Kilian, 2005). In other words, in-sample forecasts draw an unduly optimistic picture of a model's predictive power. Including out-of-sample observations will improve the predictions of a forecast model, as the power of forecast evaluation tests is strongest with long out-of-sample periods (Hansen & Timmermann, 2012). Therefore, to avoid potential overfitting issues, we will split the sample into two separate periods.

Table 1. – Data sample

Data sample	**Begin date**	**End date**	**Observations**
Full sample	01/05/2000	03/27/2018	4,078
In-sample period	01/05/2000	08/06/2004	1,000
Out-of-sample period	08/09/2004	03/27/2018	3,078

Table 1. – Brief summary of the different sample periods used in this paper.

Next, we will specify the following volatility-forecasting models in R by using the "rugarch" package: ARCH(1), GARCH(1,1), GJR-GARCH(1,1), CGARCH(1,1), APARCH(1,1), and GARCHX(1,1,1).[7] Since the order structure of these conditional variance equations offer numerous combinations, we focus only on parsimonious forecast models because of their computational efficiency and simplicity. Notice that this paper tries to find out what makes a specific GARCH model so good in volatility forecasting rather than discovering a superior model that has the best specifications by chance. Therefore, we will use only one lag order for all forecast models.

Thereafter, we will conduct rolling window direct forecasts with daily re-estimations of model parameters under Gaussian distribution to predict h-step(s) ahead out-of-sample conditional variances, or $\hat{\sigma}^2_{t+h|t}$. According to Christoffersen (2012), a general rule of thumb is to use a forecasting origin of at least 1,000 observations. Therefore, for the re-estimations of model parameters, we will use a rolling window of 1,000 observations that moves one-step forward each time. We are interested to forecast $\hat{\sigma}^2_{t+h|t}$ at different horizons to evaluate the short- and medium-term predictive ability of several GARCH models. The horizons used in this paper are one-day, one-week, one-month, and one-quarter.[8] Lastly, we will evaluate $\hat{\sigma}^2_{t+h|t}$ series by using RMSE and QLIKE, and eventually assess the significance of these evaluation tests by using DM tests and MCS.

[7] See Ghalanos (2019) for documentation of this R package.
[8] Here we assume that a week has five trading days, whereas a month and a quarter have, on average, 21 and 63 trading days, respectively, see Christoffersen (2012).

§4. Data

In this paper, we will use major equity indices from four different countries, of which two from developed markets and two from emerging markets. Therefore, we will use the S&P 500, HSI, IPC, and KOSPI indices in our research, respectively. Time series data such as index prices and realised volatility estimations are retrieved from Oxford-Man Institute's realised library. The database illustrates a variety of realised volatility estimations that are based on different sampling methods. This paper uses the 5-minute and 10-minute subsampled realised variances as proxies. Furthermore, to determine the range-based variance proxies, the high and low prices of these four equity indices are retrieved from Bloomberg. The daily log returns of equity indices are calculated as $r_t = \ln(P_t/P_{t-1})$, see Hull (2012). Lastly, daily prices of the Cboe Volatility Index, or VIX, are retrieved from Bloomberg as well. The sample period of this novel dataset starts from January 5, 2000, to March 27, 2018, which yields 4,078 daily observations in total, as Oxford-Man Institute's database starts from January 2000 onwards. Importantly, due to market holidays and possible mistakes in data entry, all rows that contain at least one missing value are left out from the sample. In other words, if a certain equity index has a missing value at time $t + k$, then the corresponding row including the non-missing values of other equity indices will be omitted.

Table 2 illustrates the general information of the equity indices that are used in this paper. The New York Stock Exchange (NYSE) and Stock Exchange of Hong Kong (SEHK) are considered developed markets, whereas Bolsa Mexicana de Valores (BMV) and Korea Exchange (KRX) are labelled as emerging markets (MSCI, 2018).

Table 2. – List of stock indices

Country	Stock exchange	Major stock index	Constituents
Hong Kong	Stock Exchange of Hong Kong	Hang Seng Index (HSI)	50
Mexico	Bolsa Mexicana de Valores	Índice de Precios y Cotizaciones (IPC)	35
South Korea	Korea Exchange	Korea Composite Stock Price Index (KOSPI)	751
United States	New York Stock Exchange	Standard & Poor's 500 (S&P 500)	500

Table 2. – The first, second, third, and fourth columns display the country, stock exchange, stock index, and number of constituents, respectively. The New York Stock Exchange and Stock Exchange of Hong Kong are considered developed markets, while the remaining two are labelled as emerging markets according to MSCI (2018).

In the next subsection, descriptive statistics of the sample data will be explained in more detail. Additional descriptions of certain time series and their corresponding tables and graphs can be found in the **Appendix** sections.

4.1. Descriptive statistics

Table 3 displays summary statistics of the daily logarithmic returns of these stock indices. We observe that return distributions of these stock indices tend to follow a leptokurtic distribution since the kurtosis estimates are all greater than three. Further, the average logarithmic returns and standard deviations of these indices tend to be close to zero and one, respectively. In addition, the Jarque-Bera test statistics show that normality assumption of the sample data is rejected at a 1% significance level in every case, which implies that the sample stock returns are likely to follow a non-normal distribution.[9] In addition, **Figure 1** illustrates the kernel density plots of return distributions of these equity indices. These kernel density plots confirm the leptokurtic shape of the return distributions, which is not surprising because stock returns tend to have fat tails (see **Section 2.2.1**). Note that the return distribution of the KOSPI (see **Panel D**) is slightly skewed to the left compared to other indices.

Table 3. – Summary statistics of stock index returns

Indices	Mean (%)	Std. dev. (%)	Min. (%)	Max. (%)	Skewness	Kurtosis	Jarque-Bera
S&P 500	0.015	1.280	-9.930	10.642	-0.279	11.404	12,054
HSI	0.014	1.584	-14.695	13.407	-0.392	12.625	15,846
IPC	0.048	1.380	-12.126	10.441	-0.176	10.251	8,955
KOSPI	0.021	1.606	-16.935	11.245	-0.739	11.613	12,976

Table 3. – This table illustrates the summary statistics of the daily logarithmic returns of various stock indices. The sample period of these stock index returns starts from January 5, 2000, to March 27, 2018, which in turn gives 4,078 observations of daily trading returns. Note that the above Jarque-Bera test statistics reject the null hypothesis, which states that the sample data is normally distributed, in each case at a significance level of 1%.

The correlations of the four equity indices returns are demonstrated in **Table 4**. From the table, we observe that returns of equity indices from the same continent tend to be stronger correlated with each other than equity indices from other continents. For instance, S&P 500 has a positive correlation coefficient of 0.69 with IPC, while the correlations with HSI and KOSPI are significantly weaker. One could suggest using dynamic conditional correlations but that is beyond the scope of this paper.

Table 4. – Correlations of stock indices returns

Indices	HSI	IPC	KOSPI	S&P 500
HSI	1.00			
IPC	0.33	1.00		
KOSPI	0.63	0.30	1.00	
S&P 500	0.27	0.69	0.22	1.00

Table 4. – This matrix displays the correlation coefficients of equity indices returns. The data sample period starts from January 5, 2000, to March 27, 2018, which contains 4,078 observations in total.

[9] Further explanation about the Jarque-Bera normality test is provided in **Appendix Section A.1**.

Figure 1. – Kernel density plots of stock index returns

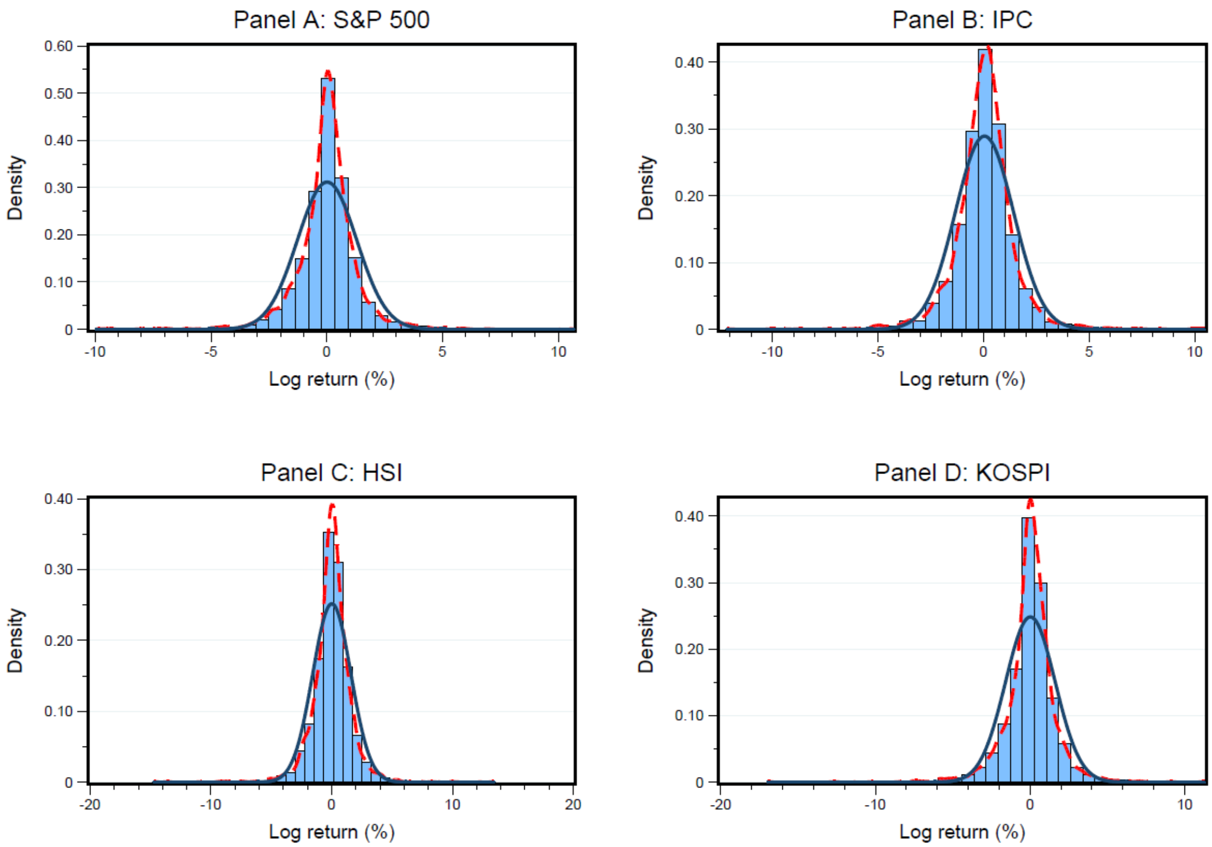

Figure 1. – This graph illustrates the return distribution of the S&P 500, IPC, HSI, and KOSPI indices. In each panel, the navy-blue curve illustrates the standard normal distribution meanwhile the red-coloured dashes display the kernel density. The sample period of these stock returns starts from January 5, 2000, to March 27, 2018, which contains 4,078 observations.

Furthermore, **Table 5** illustrates the summary statistics of the realised and range-based volatility proxies. The distributions of both *RV* and *RP* tend to show a non-normal shape as the skewness and kurtosis figures are significantly large. This also confirmed by the significantly large Jarque-Bera test statistics. Notice that the standard deviation of 5-minute and 10-minute subsampled realised volatility series of the S&P 500 is substantially larger compared to other indices in **Panels A** and **B**, meanwhile in **Panels C** and **D** the two range-based variance proxies of the KOSPI seem to have the largest standard deviations. Besides, it seems that the outliers of range-based variance proxies of these equity indices tend to be relatively more dispersed than the realised variance ones.

Table 5. – Summary statistics of volatility proxies

Indices	Mean (%)	Std. dev. (%)	Min. (%)	Max. (%)	Skewness	Kurtosis	Jarque-Bera
Panel A: *RV* (5-min)							
S&P 500	1.118	2.583	0.012	77.477	11.671	239.684	9,611,194
HSI	1.019	1.775	0.048	43.730	10.280	172.901	4,976,700
IPC	0.871	1.890	0.044	52.248	12.491	249.963	10,469,385
KOSPI	1.271	2.256	0.057	59.437	9.297	161.995	4,354,140
Panel B: *RV* (10-min)							
S&P 500	1.139	2.616	0.013	77.841	11.228	228.166	8,700,413
HSI	1.018	1.832	0.038	55.101	12.163	259.519	11,281,405
IPC	1.016	2.093	0.041	45.485	9.757	144.092	3,447,227
KOSPI	1.278	2.266	0.041	62.899	9.674	181.920	5,503,042
Panel C: *RP* (Eq. 2)							
S&P 500	1.030	2.490	0.008	42.884	9.164	115.932	2,224,133
HSI	1.010	2.679	0.029	112.325	22.467	799.842	108,232,879
IPC	1.181	2.220	0.021	43.812	7.593	95.862	1,504,436
KOSPI	1.275	2.880	0	90.512	13.316	308.202	15,947,958
Panel D: *RP* (Eq. 3)							
S&P 500	0.912	2.272	0.008	59.205	11.526	199.766	6,668,932
HSI	0.964	2.440	0.030	103.590	22.301	810.743	111,199,939
IPC	1.014	1.841	0.023	42.914	8.171	113.806	2,131,610
KOSPI	1.235	2.891	-0.432	102.172	15.047	410.165	28,323,237

Table 5. – This table illustrates the summary statistics of different volatility proxies, namely the realised and range-based variances. The sample period starts from January 5, 2000, to March 27, 2018, which results in 4,078 daily observations for all proxies. Note that the Jarque-Bera test statistics reject the null hypothesis in every case at a significance level of 1%.

Since the aim of this paper is to find out the properties of different forecast models, we are highly interested in the underlying volatility of these equity indices. For the sake of brevity, figures of S&P 500 will be demonstrated in this subsection. **Figure 2** illustrates the annualised volatility of the S&P 500 index on a daily basis over the period January 5, 2000, to March 27, 2018. The computations of annualised volatilities are done similarly as in Patton (2011a) using $\sigma_t^A = \sqrt{252 \times \widehat{VAR}_t}$, where $\widehat{VAR}_t$ is the variance proxy at time t. The figure illustrates volatility spikes that are scattered over the sample period. However, the magnitude of these volatility spikes varies among the four volatility proxies at certain moments. For instance, **Panels A** and **B** tend to display larger volatility spikes during the financial crisis of 2008 compared to **Panels C** and **D**. Additionally, the panels illustrate the clustering of volatility in certain periods as well. It seems that sizable changes in variance tend to be followed by relatively large changes, while small changes in variance tend to be followed by relatively small changes.[10] Similar results for volatility clustering can be found for other equity indices as well. The

[10] Mandelbrot (1963) used equivalent words to outline the observation of volatility clustering of price variation. His exact words were: *"…large changes tend to be followed by large changes – of either sign – and small changes tend to be followed by small changes…"* (p. 418).

corresponding line plots of these equity indices are reported in **Appendix Section C** (see **Figure C1**, **Figure C2**, and **Figure C3**).

Figure 2. – Annualised volatility proxies of S&P 500 index

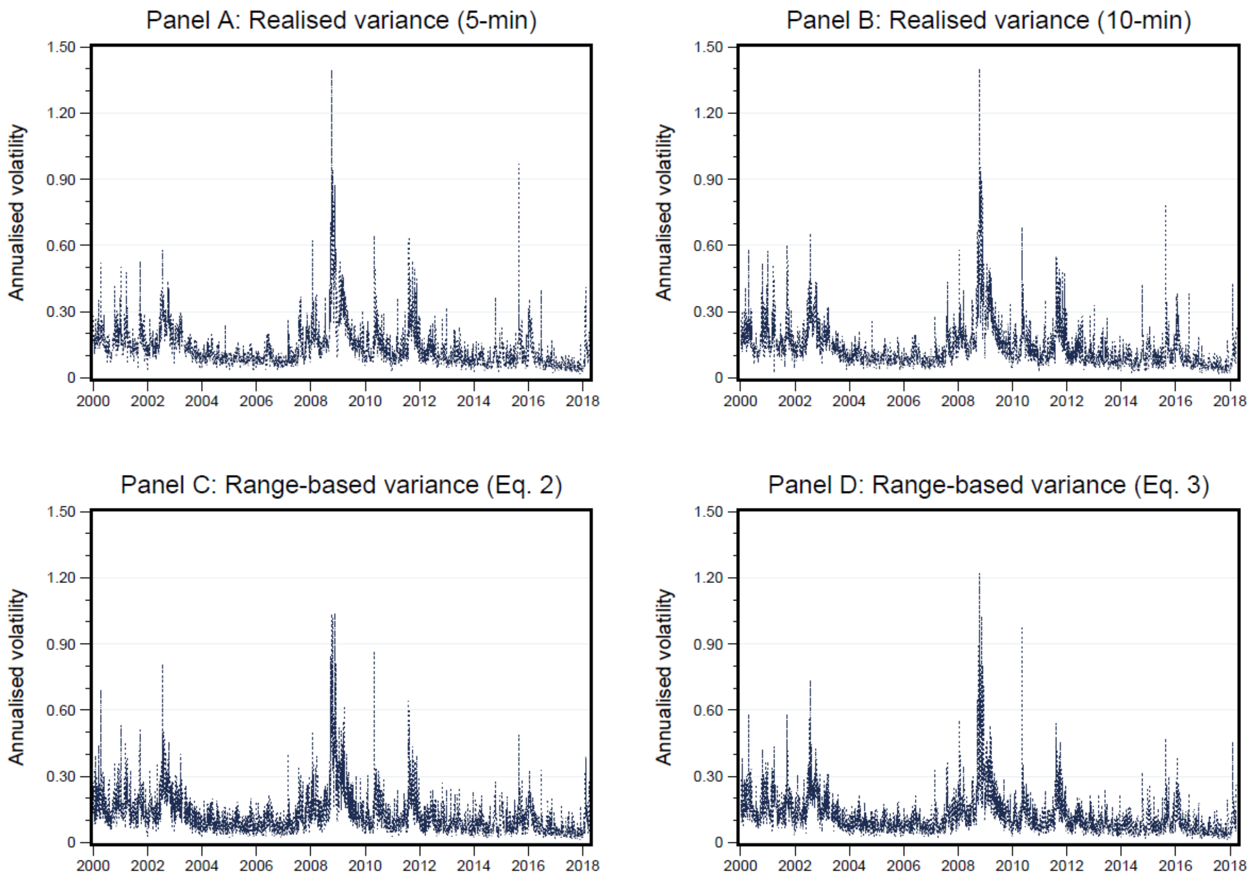

Figure 2. – This figure displays volatility estimates of the S&P 500 index based on four different proxies, namely the 5-minute and 10-minute *RV*, and the two range-based variances based on **Eq. (2)** and **Eq. (3)**. The volatility estimates are annualised, using the following formula from Patton (2011a): $\sigma_t^A = \sqrt{252 \times \widehat{VAR}_t}$. Furthermore, the vertical axes in all panels are scaled up by 1/100. The sample period starts from January 5, 2000, to March 27, 2018, which contains 4,078 observations.

In addition, **Figure 3** illustrates the autocorrelation plots, or correlograms, of the four variance estimators of the S&P 500. We observe a significant decay in the autocorrelation coefficients of the realised variance and range-based variance proxies after dozens of lags. After a while, the autocorrelation of these volatility proxies tends to converge to zero in all four panels. Similar patterns can be found for the other equity indices as well; see **Figure C4**, **Figure C5**, and **Figure C6** in the **Appendix**. These correlograms suggest that asset return volatility of the S&P 500, HSI, IPC, and KOSPI indices tends to be persistent, given the relatively high autocorrelation values in the panels below. Hence, the existence of long-memory property for these equity indices is probable.

Figure 3. – Autocorrelations of S&P 500 volatility proxies

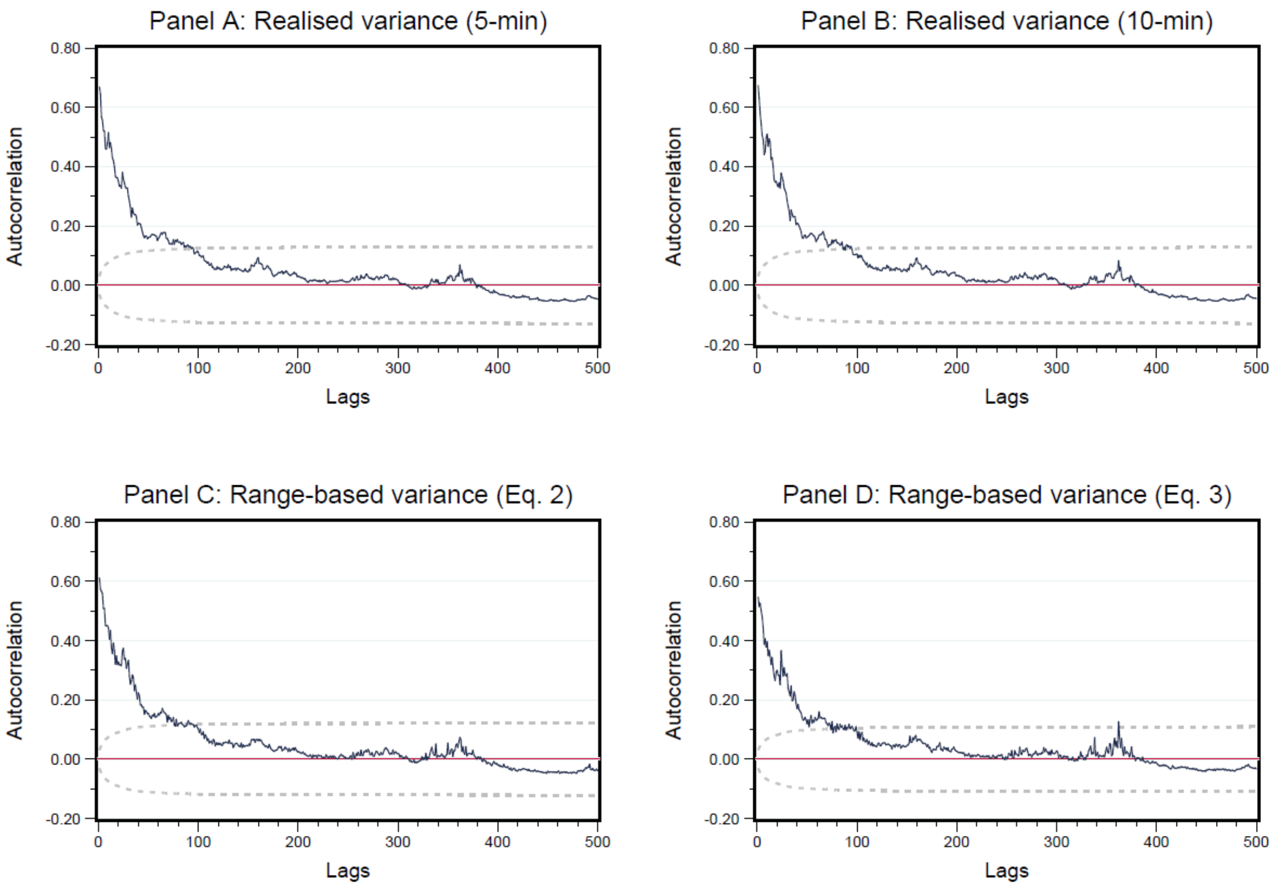

Figure 3. – In this figure, correlograms of the four different volatility proxies of the S&P 500 are demonstrated in the above panels. A window of 500 lags is used in all four panels. Furthermore, the grey dashes represent the Bartlett's formula for $MA(q)$ 95% confidence bands, i.e. 5% significance limits.

We find the partial autocorrelation plots of the four different variance estimators of the S&P 500 index in **Figure 4**. Again, we observe clear signs of serial correlation from the graph panels, but compared to the previous figure the partial autocorrelation function illustrates a relatively fast decaying pattern in these four panels. The serial correlation coefficients of these volatility proxy series converge to zero just after several lags and remain relatively stable thereafter. Similar patterns can be found for other equity indices as well; see **Figure C7**, **Figure C8**, and **Figure C9** in the **Appendix**. Compared to the previous correlograms, the partial autocorrelation plots suggest that persistence in asset return volatility does not last very long.

Figure 4. – Partial autocorrelations of S&P 500 volatility proxies

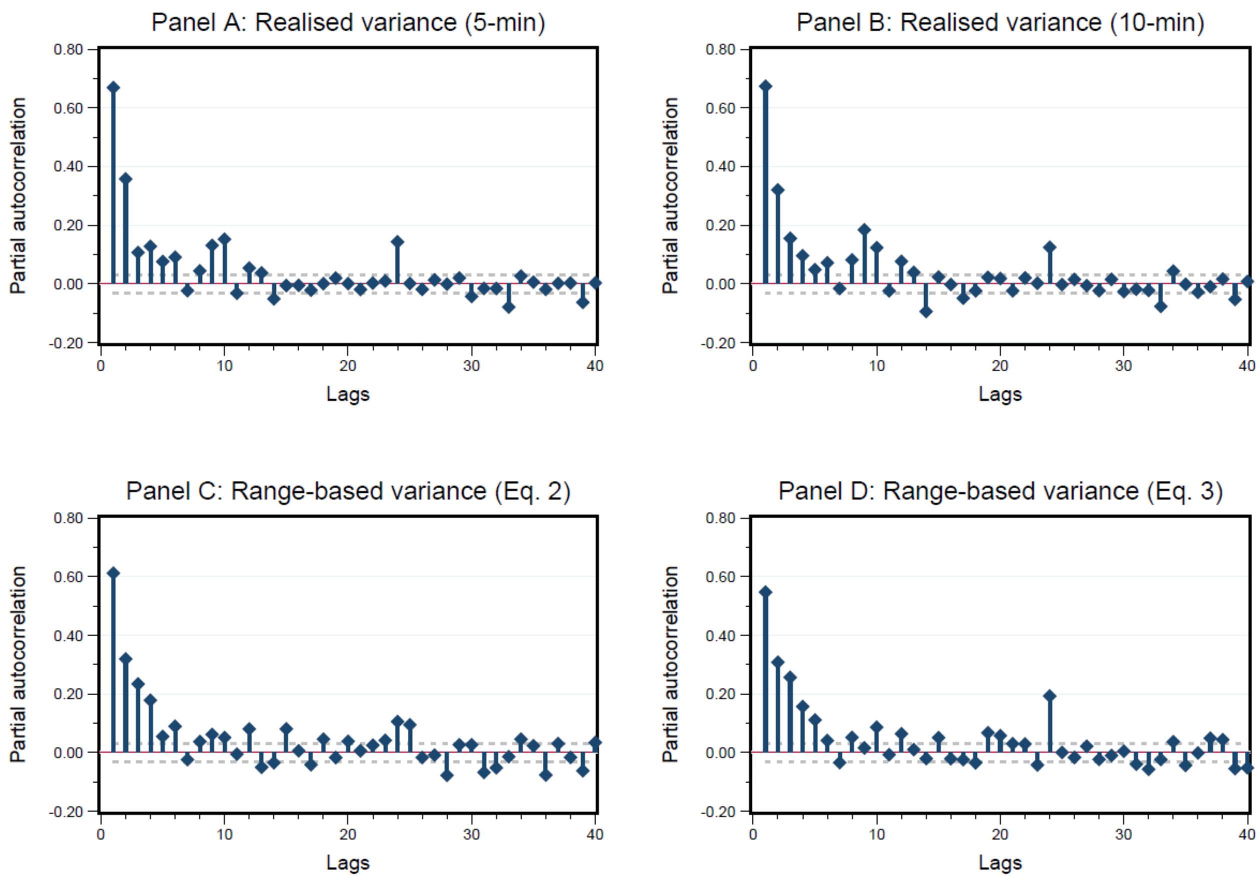

Figure 4. – In this figure, partial autocorrelations of the four different volatility proxies of the S&P 500 are demonstrated in the above panels. A window of 40 lags is used in all four panels. Furthermore, the grey dashes represent the 95% confidence bands, which can be calculated as $se = 1/\sqrt{n}$.

Moreover, to check quantitatively the occurrence of serial correlation in time series, we can use the Ljung-Box Q-test.[11] The results are demonstrated in **Table A3** in the **Appendix**. In short, based on these test results, there is sufficient statistical evidence to conclude that these variance estimators exhibit serial correlation. As an alternative to the Ljung-Box Q-test, this paper also uses Engle's ARCH test to check for serial dependence in the squared residual series.[12] In addition, **Table A4** illustrates the results of Engle's ARCH test. We find sufficient statistical evidence to conclude that the squared errors of the equity indices returns tend to be serially correlated.

[11] Further description regarding the Ljung-Box test is provided in **Appendix Section A.3**.
[12] Engle's ARCH test is described in more detail in **Appendix Section A.4**.

Figure 6, on the next page, illustrates the index price level, logarithmic returns, and squared returns of the S&P 500, which are displayed in **Panels A**, **B**, and **C**, respectively. Interestingly, it seems that a downward trend in the index price causes significant spikes in the daily returns, which tend to concur with the considerable jumps in the volatility series, or squared return in this case. Notice that squared return is a noisy proxy measure of volatility. Yet, squared return does stress the dispersion of asset returns in a convenient way. This graphical finding seems to coincide with the notion of leverage effect, which is explained earlier in **Section 2.2.4**. Similar results can be found for the other equity indices as well; see **Figure C10, Figure C11**, and **Figure C12** in the **Appendix**. Additionally, **Figure 5** displays the historical price pattern of the VIX. It seems that the VIX Index has the tendency to follow the opposite trend relative to the S&P 500. The VIX can be used as a variance proxy as well, which is shown in **Figure C13** from the **Appendix**. We observe that the VIX variance proxy closely tracks the realised variance and range-based variance proxies.

Figure 5. – Historical behaviour of the VIX Index

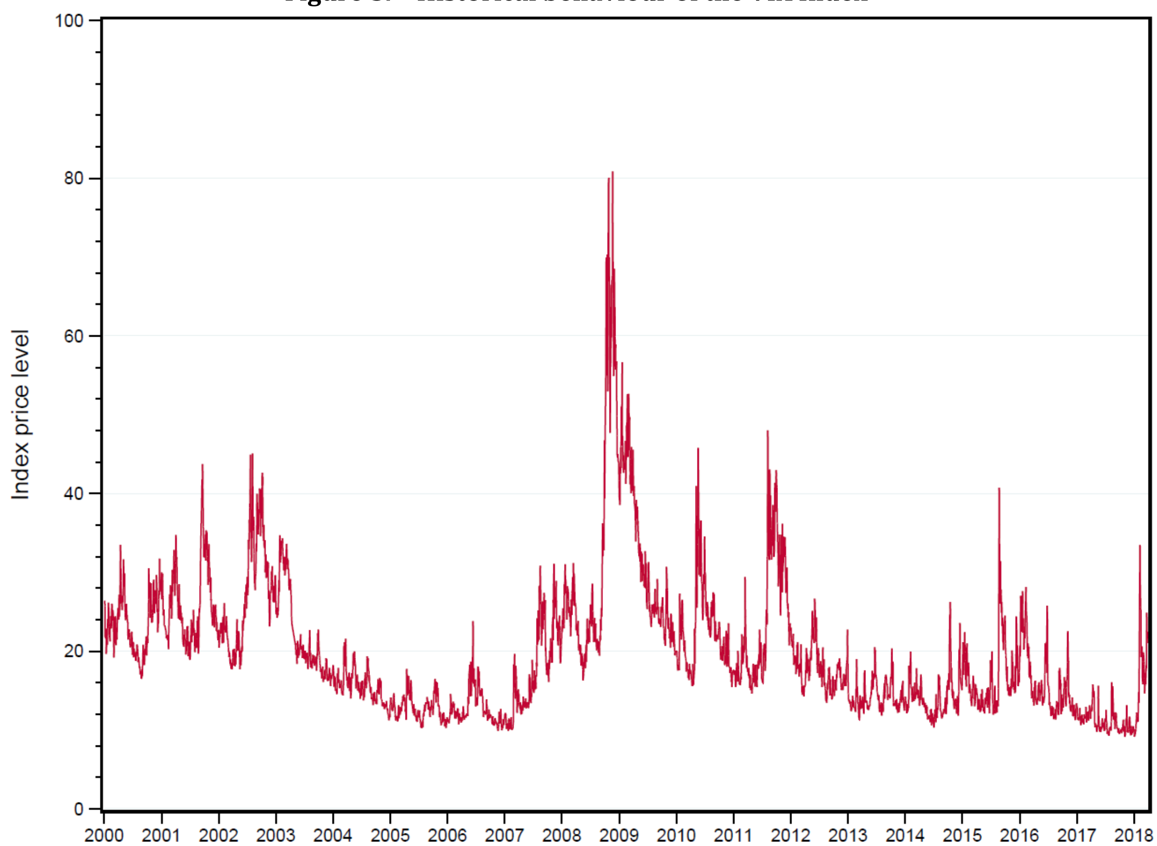

Figure 5. – Cboe Volatility Index, or VIX. The sample period starts from January 5, 2000, to March 27, 2018, which contains 4,078 observations. This results in a sample mean price of $19.83 with minimum and maximum prices of $9.14 and $80.86, respectively, and a standard deviation of $8.80.

Figure 6. – S&P 500 index

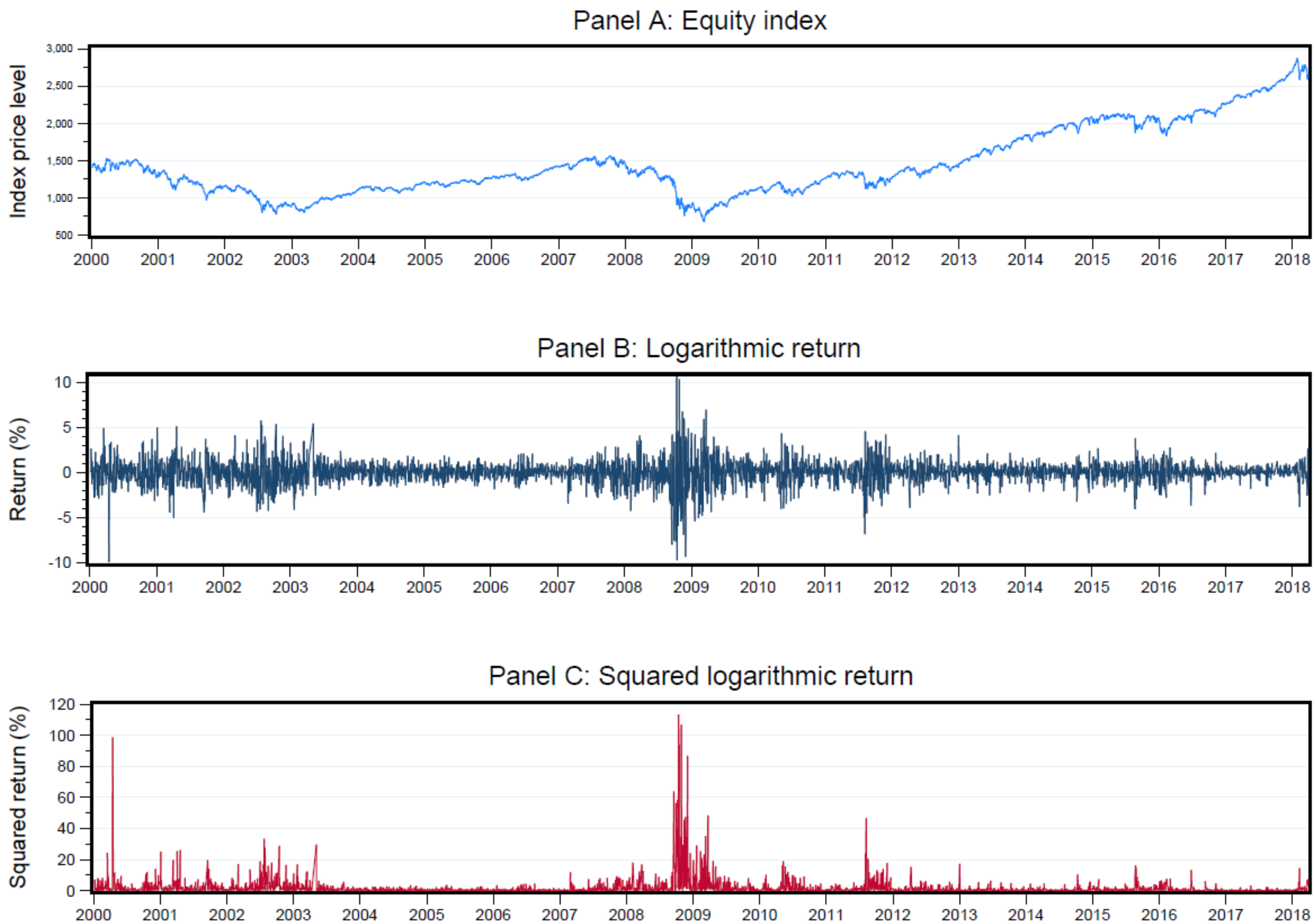

Figure 6. – This figure illustrates the historical data of the S&P 500 index price level, log returns, and squared returns. These are displayed in Panels A, B, and C, respectively. The sample period starts from January 5, 2000, to March 27, 2018, which contains 4,078 observations.

§5. Empirical results

5.1. In-sample fitting results

Table 6 illustrates the in-sample fitting results of the GARCH-type models. Based on the goodness-of-fit tests, the overall results seem to be mixed as there is no single model that fits the equity indices best in all cases. We observe that GARCHX(1,1,1) is the best-fitting model in our subset of volatility-forecasting models for S&P 500, meanwhile the APARCH(1,1) model fits the in-sample data of IPC and KOSPI best. For HSI, the APARCH(1,1) model is preferred under the log-likelihood criteria but the information criteria tests suggest that GJR-GARCH(1,1) is a better fitting model.

Table 6. – Results of in-sample goodness-of-fit tests

Forecast model	Log-likelihood	AIC	BIC
Panel A: S&P 500			
ARCH(1)	-1,768.47	3.543	3.558
GARCH(1,1)	2,918.10	-5.828	-5.809
GJR-GARCH(1,1)	2,930.49	-5.851	-5.826
CGARCH(1,1)	2,917.74	-5.824	-5.794
APARCH(1,1)	2,936.22	-5.860	-5.831
GARCHX(1,1,1)	2,951.96***	-5.894***	-5.869***
Panel B: IPC			
ARCH(1)	2,687.13	-5.368	-5.354
GARCH(1,1)	2,757.79	-5.508	-5.488
GJR-GARCH(1,1)	2,768.45	-5.527	-5.502
CGARCH(1,1)	2,772.69	-5.533	-5.504
APARCH(1,1)	2,776.79***	-5.542***	-5.512***
GARCHX(1,1,1)	2,759.08	-5.508	-5.484
Panel C: HSI			
ARCH(1)	2,726.03	-5.446	-5.431
GARCH(1,1)	2,764.30	-5.521	-5.501
GJR-GARCH(1,1)	2,786.66	-5.563***	-5.539***
CGARCH(1,1)	2,773.50	-5.535	-5.506
APARCH(1,1)	2,786.91***	-5.562	-5.532
GARCHX(1,1,1)	2,765.63	-5.521	-5.497
Panel D: KOSPI			
ARCH(1)	2,362.99	-4.720	-4.705
GARCH(1,1)	2,387.86	-4.768	-4.748
GJR-GARCH(1,1)	2,401.64	-4.793	-4.769
CGARCH(1,1)	2,400.67	-4.789	-4.760
APARCH(1,1)	2,408.77***	-4.806***	-4.776***
GARCHX(1,1,1)	2,403.11	-4.796	-4.772

Table 6. – This table illustrates the results of in-sample goodness-of-fit tests for the above-mentioned volatility-forecasting models. The in-sample period starts from January 5, 2000, to August 6, 2004, resulting in 1,000 observations. *** indicates the most preferred forecasting model in terms of in-sample fit under the log-likelihood, AIC, and BIC criteria.

Furthermore, **Table B1** from the **Appendix** illustrates the in-sample estimations of model parameters of all forecasting models. The results show mixed results for some GARCH-type models. For instance, we observe that the APARCH(1,1) model has several statistically significant parameters for all equity indices, meanwhile the GARCHX(1,1,1) model has in some cases only a few statistically significant parameters. The former result can possibly be explained by the model's flexibility that allows for many forms. It is important to mention that these in-sample estimations do not say anything about predictive ability of our GARCH-type models.

5.2. Forecast evaluation

This subsection discusses the direct forecast results of the GARCH-type models at multiple horizons. Again, only tables and figures related to the results of S&P 500 will be presented here for the sake of brevity. The rest can be found in the **Appendix** sections.

• ***5.2.1. Results of loss functions***

Table 7 illustrates the forecast loss coefficients of the GARCH-type models of S&P 500 at multiple horizons. Similar to Brownlees et al. (2011), when h increases, the loss coefficients tend to show larger discrepancies in forecast errors due to the uncertainty over longer horizons. From all the forecasting models, we clearly observe that the GARCHX(1,1,1) model has the lowest loss coefficients in all cases, suggesting that this model predicts relatively more accurate forecasts of both short- and medium-term conditional variances for the S&P 500 index than other models in this sample. Reason why the GARCHX model performs relatively well here has probably to do with the inclusion of the VIX variance proxy, since it tracks the *RV* and *RP* proxies relatively well (see **Figure C13** in the **Appendix**), resulting in sharper volatility forecasts and thus smaller discrepancies in forecast errors. In contrast, the forecast results of the ARCH(1) model are relatively weak compared to other models in all cases. Yet, it is no surprise that the ARCH(1) model underperforms here, since other forecast models are extensions of the ARCH model, see **Section 3.2**.

Moreover, APARCH(1,1) seems to be a good alternative, as in second best model, for GARCHX(1,1,1) when it comes to one-step ahead forecasts, as both RMSE and QLIKE coefficients of APARCH(1,1) are second lowest for all variance proxies in this sample. However, for $h > 1$, other forecasting models are preferred as alternative for the GARCHX(1,1,1) model, due to their relatively small loss coefficients. For instance, the CGARCH(1,1) model has the second lowest RMSE coefficients for all variance proxies. Meanwhile, QLIKE ranks the ordinary GARCH(1,1) model as second best under both the *RV* and *RP* criteria for $h = 21$ and $h = 63$. For one-week ahead forecasts, QLIKE ranks APARCH(1,1) and GJR-GARCH(1,1) as second best based on *RV* proxies and *RP* proxies, respectively.

Figure 7 and **Figure 8** (see pages 34 and 35) illustrate the one-step ahead conditional variance forecasts of the S&P 500. Notice that the red and blue lines display the in-sample forecasts and out-of-sample forecasts, respectively. The ARCH(1) forecasts are substantial in size compared to other

GARCH models. In contrast to other GARCH forecasts, the ARCH(1) series seems to be less refined as well, which is not surprising since the ARCH model is only a linear function of past squared errors. In addition, we observe that the GARCHX(1,1,1) forecasts tend to be relatively small compared to the other GARCH models in this sample.

Table 7. – Forecast loss coefficients of S&P 500

	$h = 1$		$h = 5$		$h = 21$		$h = 63$	
Forecast model	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**
Panel A: *RV* (5-min)								
ARCH(1)	6.280	1.139	6.617	1.139	6.807	1.258	6.963	1.481
GARCH(1,1)	2.238	0.352	2.568	0.437	3.143	0.601	3.648	0.754
GJR-GARCH(1,1)	2.062	0.319	2.534	0.421	3.181	0.623	3.728	0.770
CGARCH(1,1)	2.170	0.334	2.477	0.429	3.014	0.635	3.495	0.775
APARCH(1,1)	2.045	0.315	2.498	0.416	3.162	0.677	3.708	0.772
GARCHX(1,1,1)	1.915***	0.267***	2.281***	0.375***	2.705***	0.567***	3.047***	0.720***
Panel B: *RV* (10-min)								
ARCH(1)	6.262	1.151	6.616	1.156	6.798	1.279	6.958	1.475
GARCH(1,1)	2.238	0.371	2.547	0.452	3.133	0.613	3.638	0.761
GJR-GARCH(1,1)	2.069	0.338	2.508	0.433	3.175	0.638	3.720	0.775
CGARCH(1,1)	2.166	0.353	2.453	0.443	3.002	0.648	3.485	0.780
APARCH(1,1)	2.045	0.334	2.471	0.431	3.156	0.694	3.700	0.776
GARCHX(1,1,1)	1.907***	0.286***	2.265***	0.387***	2.700***	0.578***	3.031***	0.726***
Panel C: *RP* (Eq. 2)								
ARCH(1)	6.529	1.270	6.596	1.319	6.788	1.520	6.925	1.593
GARCH(1,1)	2.313	0.539	2.557	0.601	3.053	0.749	3.585	0.877
GJR-GARCH(1,1)	2.177	0.505	2.529	0.577	3.103	0.769	3.673	0.886
CGARCH(1,1)	2.227	0.522	2.436	0.589	2.914	0.780	3.422	0.888
APARCH(1,1)	2.148	0.502	2.486	0.578	3.083	0.833	3.653	0.896
GARCHX(1,1,1)	1.864***	0.438***	2.171***	0.528***	2.566***	0.701***	2.926***	0.832***
Panel D: *RP* (Eq. 3)								
ARCH(1)	6.438	1.270	6.630	1.286	6.758	1.379	6.907	1.560
GARCH(1,1)	2.380	0.535	2.645	0.587	3.071	0.723	3.521	0.827
GJR-GARCH(1,1)	2.274	0.502	2.627	0.564	3.129	0.737	3.604	0.828
CGARCH(1,1)	2.258	0.513	2.509	0.572	2.918	0.752	3.347	0.830
APARCH(1,1)	2.245	0.501	2.591	0.568	3.109	0.798	3.583	0.836
GARCHX(1,1,1)	1.843***	0.437***	2.137***	0.514***	2.485***	0.661***	2.788***	0.770***

Table 7. – This table displays the empirical loss coefficients of various GARCH-type models of the S&P 500 index at multiple horizons under different variance proxies. For the one-step, five-steps, 21-steps, and 63-steps ahead direct forecasts, we obtained 3,078, 3,074, 3,058, and 3,016 out-of-sample observations, respectively. Note that RMSE coefficients have been scaled up for convenience sake. *** indicates the preferred forecasting model that yields the lowest loss function coefficient, and thus illustrates the best out-of-sample forecast performance in this sample.

In contrast to the S&P 500 predictions, the forecast results of HSI, IPC, and KOSPI display a slightly different picture (see **Table B2**, **Table B3**, and **Table B4** from the **Appendix)**. To make the results more concise and accessible for the reader, **Table 8** summarises the forecast results of the above-mentioned tables in which only the outperforming models are demonstrated for each horizon. Purely observing the loss coefficients of these tables, we conclude that CGARCH(1,1) tends to outperform GARCHX(1,1,1) in most cases, whilst APARCH(1,1) is superior in only a few cases.

These outcomes could possibly be explained by the VIX component in GARCHX, which is closely linked with the implied volatilities of the S&P 500. This covariate seems to explain the dispersion of S&P 500 returns more accurately than the asset return volatilities of other equity indices, and therefore underperforming models like CGARCH and APARCH in some cases. **Table B5** confirms this conjecture by showing the adjusted R^2 of the out-of-sample regressions of various volatility proxies of the indices on the VIX variance proxy, suggesting that the VIX component explains a larger proportion of the variability in S&P 500 variance proxies than the ones of HSI, IPC, and KOSPI.

Table 8. – Summary of forecast models with the lowest loss coefficients

	$h = 1$		$h = 5$		$h = 21$		$h = 63$	
Forecast model	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**
Panel A: *RV* (5-min)								
S&P 500	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX
HSI	CGARCH	GARCHX	CGARCH	GARCHX	CGARCH	CGARCH	CGARCH	GARCHX
IPC	APARCH	CGARCH	CGARCH	APARCH	CGARCH	CGARCH	CGARCH	APARCH
KOSPI	APARCH	CGARCH	APARCH	CGARCH	CGARCH	CGARCH	CGARCH	CGARCH
Panel B: *RV* (10-min)								
S&P 500	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX
HSI	CGARCH	GARCHX	CGARCH	GARCHX	CGARCH	CGARCH	CGARCH	GARCHX
IPC	APARCH	CGARCH	CGARCH	CGARCH	CGARCH	CGARCH	CGARCH	APARCH
KOSPI	APARCH	CGARCH	APARCH	CGARCH	CGARCH	CGARCH	CGARCH	CGARCH
Panel C: *RP* (Eq. 2)								
S&P 500	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX
HSI	CGARCH	GARCHX	CGARCH	GARCHX	CGARCH	CGARCH	CGARCH	GARCHX
IPC	APARCH	APARCH	CGARCH	APARCH	CGARCH	CGARCH	CGARCH	APARCH
KOSPI	APARCH	CGARCH	APARCH	CGARCH	CGARCH	CGARCH	CGARCH	CGARCH
Panel D: *RP* (Eq. 3)								
S&P 500	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX	GARCHX
HSI	CGARCH	GARCHX	CGARCH	GARCHX	CGARCH	CGARCH	CGARCH	GARCHX
IPC	CGARCH	CGARCH	CGARCH	CGARCH	CGARCH	CGARCH	CGARCH	APARCH
KOSPI	APARCH	CGARCH	APARCH	CGARCH	CGARCH	CGARCH	CGARCH	CGARCH
Panel E: Statistics								
CGARCH	5	7	8	6	12	12	12	4
APARCH	7	1	4	2	–	–	–	4
GARCHX	4	8	4	8	4	4	4	8

Table 8. – This table summarises the forecast results of **Table 7**, **Table B2**, **Table B3**, and **Table B4** by showing only the models with the lowest RMSE and QLIKE losses. Further, Panel E counts the number of times a model is mentioned in the corresponding column. CGARCH, GARCHX, and APARCH models are mentioned 66, 44, and 18 times in total, respectively. By further decomposing these statistics into four horizons, we gain additional insights. For one-step ahead, CGARCH, GARCHX, and APARCH are mentioned 12, 12, and 8 times, respectively. For one-week ahead, CGARCH, GARCHX, and APARCH are mentioned 14, 12, and 6 times, respectively. For one-month ahead, CGARCH and GARCHX are mentioned 24 and 8 times, respectively. Lastly, for one-quarter ahead, CGARCH, GARCHX, and APARCH are mentioned 16, 12, and 4 times, respectively.

Besides comparing forecast performances of models, it is also interesting to examine the average losses across different equity indices at multiple horizons, because it may shed some light on the dissimilarities of volatility forecasts between developed markets and emerging markets. Hence, **Table 9** illustrates the average forecast losses of the four indices for each horizon and variance proxy. We observe that the mean forecast losses of S&P 500 tend to be lower than other indices for all cases under RMSE, whereas the Hang Seng Index tends to have the highest mean forecast loss compared to the rest. Nonetheless, the QLIKE-based mean losses are mixed for each horizon. For instance, indices from emerging markets tend to have the lowest mean forecast losses at longer horizons, but only occasionally at short horizons. Meanwhile, the S&P 500 has the lowest mean losses only at $h = 1$ and $h = 5$ in **Panels A** and **B** under QLIKE. However, the QLIKE-based mean losses of S&P 500 deteriorate significantly at longer horizons. Ultimately, the results below are rather ambiguous, since it is not apparent whether volatility forecasts of developed markets indices are likely to have lower losses compared to emerging markets indices, or vice versa. Therefore, further research may provide clearer picture on this, but that is beyond the scope of this paper.

Table 9. – Average forecast losses

	$h = 1$		$h = 5$		$h = 21$		$h = 63$	
Forecast model	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**
Panel A: *RV* (5-min)								
S&P 500	2.785	0.454	3.163	0.536	3.669	0.727	4.098	0.879
HSI	4.664	0.598	4.922	0.623	5.236	0.665	5.615	0.688
IPC	4.649	0.565	4.847	0.602	5.141	0.662	5.361	0.732
KOSPI	3.407	0.565	3.697	0.699	4.261	0.716	4.637	0.686
Panel B: *RV* (10-min)								
S&P 500	2.781	0.472	3.143	0.550	3.661	0.742	4.089	0.882
HSI	4.695	0.624	4.962	0.627	5.276	0.696	5.658	0.701
IPC	4.539	0.535	4.768	0.575	5.111	0.643	5.360	0.721
KOSPI	3.473	0.583	3.779	0.682	4.300	0.716	4.648	0.694
Panel C: *RP* (Eq. 2)								
S&P 500	2.876	0.629	3.129	0.699	3.585	0.892	4.031	0.995
HSI	5.052	0.771	5.334	0.812	5.677	0.913	6.088	0.880
IPC	4.536	0.571	4.721	0.611	5.066	0.686	5.347	0.765
KOSPI	3.857	0.713	4.202	0.861	4.758	0.859	5.059	0.854
Panel D: *RP* (Eq. 3)								
S&P 500	2.906	0.626	3.190	0.682	3.578	0.842	3.958	0.942
HSI	5.097	0.745	5.329	0.782	5.641	0.805	6.001	0.848
IPC	4.470	0.571	4.684	0.606	4.996	0.666	5.210	0.738
KOSPI	3.900	0.682	4.212	0.892	4.778	0.870	5.063	0.835

Table 9. – This table displays the average forecast losses across different equity indices at multiple horizons. The averages are based on the loss coefficients of GARCH-type models in **Table 7**, **Table B2**, **Table B3**, and **Table B4**.

Note that the reported forecast losses vary moderately under the 5-minute and 10-minute subsampled realised volatility proxies and the two range-based variance proxies. It is important to use several unbiased variance estimators when evaluating the validity and reliability of out-of-sample forecasts, because consistency between variance estimators leads to better understanding of a model's predictive ability. Therefore, excluding a variance proxy because of unfavourable forecast losses may potentially lead to unrepresentative conclusions.

Figure 7. – One-step ahead volatility forecasts of S&P 500

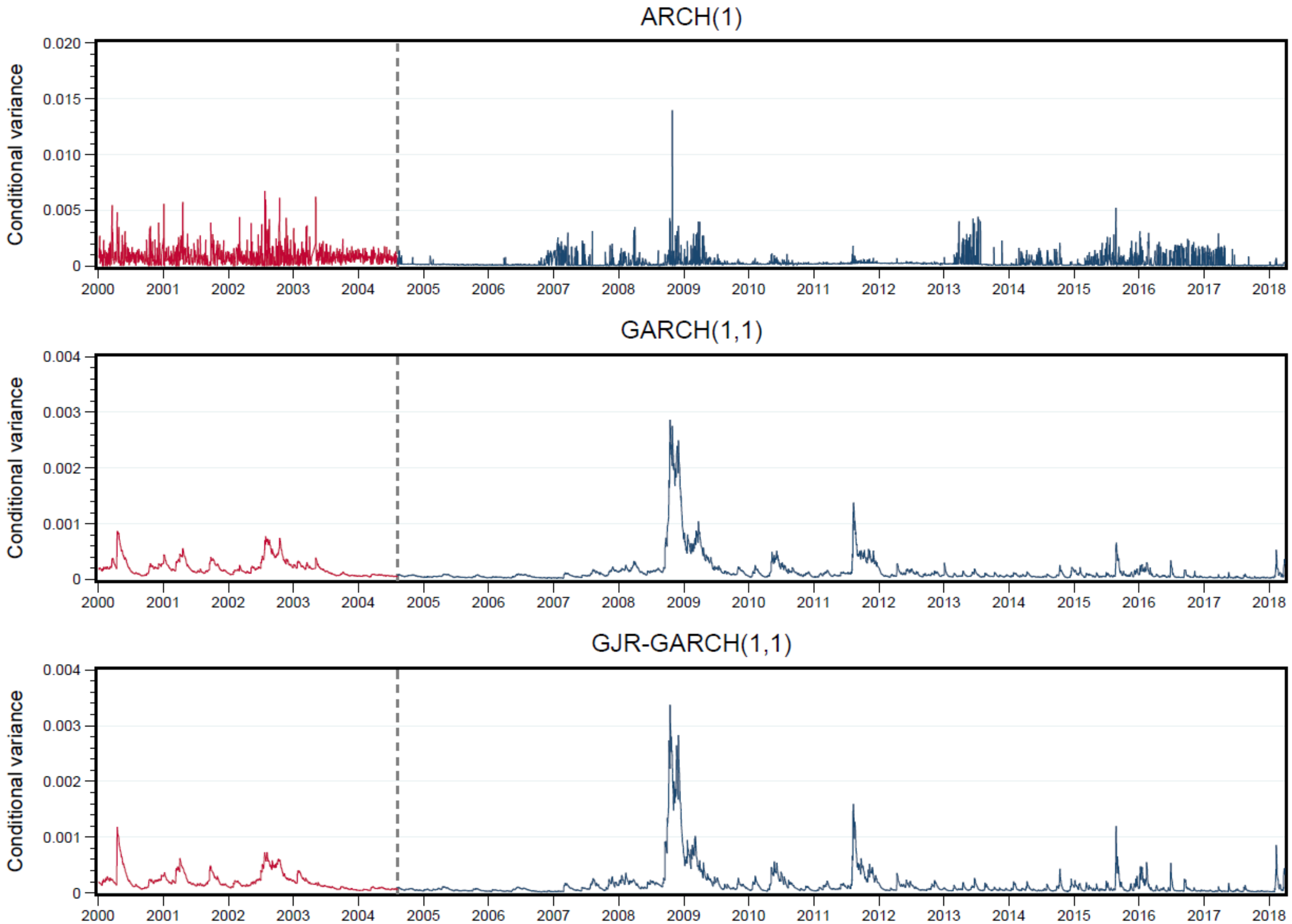

Figure 7. – This figure illustrates the rolling one-step ahead volatility forecasts of ARCH(1), GARCH(1,1), and GJR-GARCH(1,1) models. Each panel illustrates the in- and out-of-sample conditional variances, which is separated by the grey dashed line. The in-sample period starts from January 5, 2000, to August 6, 2004, resulting in 1,000 observations, whereas the forecast period starts from August 9, 2004, to March 27, 2018, resulting in 3,078 observations.

Figure 8. – One-step ahead volatility forecasts of S&P 500

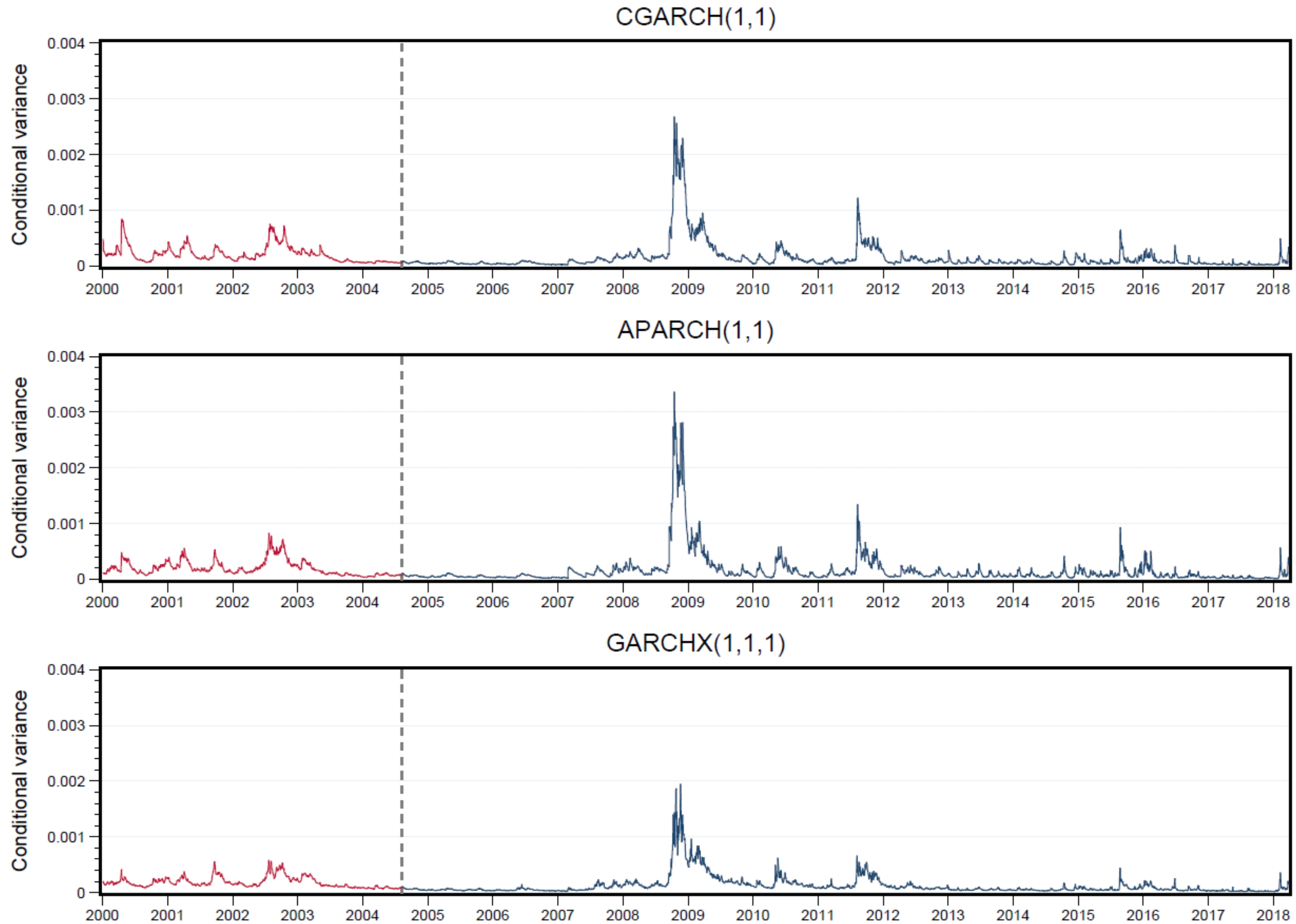

Figure 8. – This figure illustrates the rolling one-step ahead volatility forecasts of CGARCH(1), APARCH(1,1), and GARCHX(1,1,1) models. Each panel illustrates the in- and out-of-sample conditional variances, which is separated by the grey dashed line. The in-sample period starts from January 5, 2000, to August 6, 2004, resulting in 1,000 observations, whereas the forecast period starts from August 9, 2004, to March 27, 2018, resulting in 3,078 observations.

• *5.2.2. Results of Diebold-Mariano tests*

To check whether the forecast results under the RMSE and QLIKE loss functions are statistically significant, we conduct Diebold-Mariano tests for only the S&P 500 in order to perform pairwise comparisons of GARCH-type models. The DM test statistics for each horizon, which are obtained by regression of $d_{ij,t}$ on a constant with HAC standard errors, are reported in the **Appendix** section; see **Table B6**, **Table B7**, **Table B8**, and **Table B9**. From these tables, we observe positive and negative DM test statistics. A positive DM test statistic means that the column model tends to be less accurate compared to the corresponding row model, whereas a negative DM test statistic implies that the column model tends to predict more accurately than the corresponding row model.

According to these four tables, ARCH(1) seems to be the least desired model for volatility forecasting, as other GARCH models tend to generate more accurate forecasts. Further, the test statistics of medium-term horizons, $h > 5$, under QLIKE tend to be less significant than the ones of short-term horizons, suggesting that medium-term forecast outperformance of a pair of models happens mostly by chance. In other words, the two competing models are inclined to perform equally well, or bad, under QLIKE when it comes to one-month or one-quarter ahead forecasts of conditional variances of the S&P 500. This is obviously not remarkable due to the uncertain nature of long-term forecasts. What is striking is that for shorter horizons the combinations of APARCH(1,1) and GJR-GARCH(1,1), and CGARCH(1,1) and GJR-GARCH(1,1) models tend to be statistically insignificant under both RMSE and QLIKE. Note that the GJR-GARCH model can be derived from APARCH, since it is a special case of the latter. This fact may explain to some extent the insignificant DM test statistics of the APARCH/GJR-GARCH combination.

Nonetheless, by placing each forecast model in juxtaposition as a pair, we cannot effortlessly point out which forecast models are superior to other forecast models for all horizons. In fact, the Diebold-Mariano test does not provide clear answers to the overall performance of models for a specific horizon. Therefore, the use of Model Confidence Sets is more convenient in this case.

• *5.2.3. Results of MCS*

This paper uses the "MCS" package in R to determine the MCS p-values.[13] We use the block bootstrap procedure and set the parameters to $B = 1{,}000$ resamples, block length $\mathcal{L} = h$, and significance level $\alpha = 5\%$. The block length of this bootstrap procedure not only depends on the horizon of our forecast losses, but also on the potential serial correlation of these losses. Thus, the larger the forecast horizon, the larger the block length of the bootstrap. Note that we set block length $\mathcal{L} \neq h$ only for the

[13] See Catania and Bernardi (2017) for documentation of this R package.

one-step ahead forecasts, in order to take into account any possible serial correlated forecast errors. Hence, for $h = 1$, we set $\mathcal{L} = 2$. The MCS results of the S&P 500 are reported in **Table 10**, whereas the results of the remaining equity indices are illustrated in the **Appendix**, as usual. Forecast models that are included in $\widehat{\mathcal{M}}^*_{0.95}$, or 95% Model Confidence Set, are identified by asterisks.

Table 10. – Results of MCS for S&P 500

	$h = 1$		$h = 5$		$h = 21$		$h = 63$	
Forecast model	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**
Panel A: *RV* (5-min)								
ARCH(1)	0	0	0	0	0	0	0	0
GARCH(1,1)	0	0	0	0	0	0.488***	0	0.679***
GJR-GARCH(1,1)	0	0	0	0	0	0.169***	0	0.478***
CGARCH(1,1)	0	0	0	0	0	0.084***	0	0.350***
APARCH(1,1)	0	0	0	0	0	0.185***	0	0.548***
GARCHX(1,1,1)	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***
Panel B: *RV* (10-min)								
ARCH(1)	0	0	0	0	0	0	0	0
GARCH(1,1)	0	0	0	0	0	0.431***	0	0.595***
GJR-GARCH(1,1)	0	0	0	0	0	0.155***	0	0.483***
CGARCH(1,1)	0	0	0	0	0	0.091***	0	0.284***
APARCH(1,1)	0	0	0	0	0	0.149***	0	0.521***
GARCHX(1,1,1)	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***
Panel C: *RP* (Eq. 2)								
ARCH(1)	0	0	0	0	0	0	0	0
GARCH(1,1)	0	0	0	0	0	0.206***	0	0.394***
GJR-GARCH(1,1)	0	0	0	0	0	0.087***	0	0.359***
CGARCH(1,1)	0	0	0	0	0	0.107***	0	0.271***
APARCH(1,1)	0	0	0	0	0	0.084***	0	0.296***
GARCHX(1,1,1)	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***
Panel D: *RP* (Eq. 3)								
ARCH(1)	0	0	0	0	0	0	0	0
GARCH(1,1)	0	0	0	0	0	0.052***	0	0.105***
GJR-GARCH(1,1)	0	0	0	0	0	0.057***	0	0.108***
CGARCH(1,1)	0	0	0	0	0	0.115***	0	0.077***
APARCH(1,1)	0	0	0	0	0	0.050***	0	0.079***
GARCHX(1,1,1)	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***

Table 10. – This table provides the p-values of Model Confidence Sets for the S&P 500 forecast losses at multiple horizons. A block bootstrap approach of 1,000 resamples, block length equal to forecast horizon, and 5% significance level are used to determine these coefficients. Note that we set $\mathcal{L} = 2$ for the one-step ahead forecasts. *** denotes the forecast model is part of $\widehat{\mathcal{M}}^*_{0.95}$.

From the table above, GARCHX seems to be the only model included in $\widehat{\mathcal{M}}^*_{0.95}$ for all horizons under RMSE, whereas ARCH is the only model excluded from $\widehat{\mathcal{M}}^*_{0.95}$ for $h = 21$ and $h = 63$ under QLIKE. In contrast to Quaedvlieg's (2018) results, the QLIKE-based MCS for the S&P 500 expands for $h > 5$ in this paper. At first, the results of this study seem conflicting, but there may be a logical explanation for these findings. It could be that due to the likelihood of increased forecast uncertainty at longer horizons, various GARCH models are considered to be superior in this case. In other words, they tend to perform equally well for longer horizons. Altogether, the results of **Table 10** seem to suggest that the customised GARCHX model is a superior model for both short- and medium-term forecasting of conditional variances of the S&P 500. Again, it seems that adding an accurate variance proxy to the standard GARCH model improves the predictive ability in this case.

Furthermore, **Table B10** reports the MCS results related to the Hang Seng Index. Here we observe a slightly different picture in which CGARCH tends to be the sole superior model for all horizons under RMSE. Although the component GARCH is included in $\widehat{\mathcal{M}}^*_{0.95}$ for all horizons under QLIKE, the customised GARCHX model is ranked higher in terms of MCS p-value in most cases. It is remarkable that ARCH is part of the 95% MCS in **Panels B** and **C** for $h = 21$ under QLIKE, since the forecast errors of ARCH tend to be relatively large compared to other models, see previous subsubsections. Despite this, the CGARCH and GARCHX models are mostly preferred due to the ranking of MCS p-values.

Results related to the Mexican IPC index are reported in **Table B11**, where only the component GARCH model is part of $\widehat{\mathcal{M}}^*_{0.95}$ for almost all horizons under RMSE. However, when using either the 10-minute subsampled realised variance proxy or the range proxy based on **Eq. (2)**, the APARCH model is preferred for one-step ahead forecasts under RMSE. It seems that $\widehat{\mathcal{M}}^*_{0.95}$ tends to expand over the horizons under the QLIKE loss function, including forecast models such as CGARCH, APARCH, GJR-GARCH, GARCHX, and even GARCH in $\widehat{\mathcal{M}}^*_{0.95}$. Altogether, depending on the loss functions, the results suggest that CGARCH and APARCH, and sometimes GJR-GARCH and GARCHX as well, are the most preferred models for volatility forecasting, since they are ranked highest in terms of MCS p-value.

Lastly, **Table B12** illustrates the results of MCS for the Korea Composite Stock Price Index. In this case, the APARCH model tends to be the sole superior model for short horizons under RMSE, while the component GARCH model seems to be the only model in $\widehat{\mathcal{M}}^*_{0.95}$ for longer horizons under the same loss function. Based on the QLIKE loss function, almost all models are included in $\widehat{\mathcal{M}}^*_{0.95}$ across the horizons. Models that are frequently included in the 95% MCS are CGARCH, APARCH, GJR-GARCH, and GARCHX. For some horizons under QLIKE, the simple GARCH(1,1) model or even ARCH(1) are included as well. As a whole, both APARCH and CGARCH are preferred for shorter horizons, while CGARCH and GARCHX tend to be better models for long-term forecasting.

§6. Research limitations

Like all other research studies, this paper has its limitations in terms of empirical research as well. Therefore, it is important to address several of those limitations here. Firstly, the split of sample data into an in-sample and out-of-sample period is done arbitrarily in this study, i.e. without additional robustness tests. The current approach in this paper might affect the out-of-sample forecast results to a certain extent. Whether this has large implications for the validity of this empirical study is not entirely clear yet. On the other hand, finance literature has hitherto not synthesised a proper methodological guidance on the issue of selecting the right split point that separates the sample data into two parts for forecasting purposes (Hansen & Timmermann, 2012).

Secondly, this paper did not use the iterated approach next to the direct method for out-of-sample forecasts, since the former approach requires advanced programming skills. Using both methods would make a comparison between competing forecast models complete, in the sense that one should not only focus on the prediction models, but also on the forecast procedure itself. Unfortunately, the "rugarch" package in R does not have a built-in function yet for computing iterated forecasts, hence leaving it as a cumbersome programming exercise to the researcher.

Moreover, the model parameters were estimated under the classic Gaussian distribution, which is often criticised in finance literature. Mandelbrot (1963), Fama (1965), and Poon and Granger (2005) have shown evidence that the returns of financial assets tend to follow a non-normal distribution. Despite its convenience, the Gaussian distribution does not appropriately capture the heavy tails of asset return distributions (Christoffersen, 2012). As a result, it understates the probability of extreme changes in asset returns (Hull, 2015), which eventually affects the forecast accuracy of models.

Finally, this paper evaluated the direct forecast results of GARCH-type models individually at each horizon. It does not use specific evaluation tests for joint multistep testing of out-of-sample forecasts, due to lack of popular joint multi-horizon forecast evaluation tests in finance literature. Despite this, Quaedvlieg (2018) proposed an extension to the MCS approach of Hansen et al. (2011), which allows for joint multi-horizon testing of forecast models at once. A potential drawback of evaluating multi-horizon forecast performances of various models individually is that it may potentially lead to spurious outcomes, and thus resulting in faulty conclusions about the predictive ability of forecast models.

Accordingly, caution should be exercised when examining the empirical results of volatility forecasts in this research paper due to its limitations.

§7. Suggestions for further research

Throughout the years, several empirical studies have conducted research on volatility-forecasting models and their performances. Many have tried to come up with novel mathematical models and better proxy measures of market volatility to improve the predictive power of models so that more accurate trajectory of future dispersions of asset returns can be forecasted. Regardless of all the empirical studies that have been done in the past, research in volatility forecasting is still far from complete due to the complex nature of this topic. In addition, advances in the field of computer science may potentially lead to new applications and techniques for volatility-forecasting research, which might lead to new insights. Hence, further research in this field is still relevant.

For instance, one could use more complex univariate GARCH-type models, such as the Fractionally Integrated GARCH (FIGARCH) or Fractionally Integrated APARCH (FIAPARCH) models. Additionally, one could extend the GARCHX model by adding additional external regressors, such as a covariate that captures volatility-spillover effects from other financial markets (Diebold & Yilmaz, 2009), or volatility of volatility measure of an underlying equity index, e.g. Cboe VVIX Index, to pick up early signals of market volatility.[14] Furthermore, one could also evaluate forecasting performances of multivariate GARCH-type models at multiple horizons. This is interesting for investment managers who manage large multi-asset funds since they invest across a number of different asset classes, such as cash equities and FICC-related instruments. When it comes to both active and passive management of portfolios, multivariate forecast models could help asset managers with their decisions on the funds' strategic and tactical asset allocation.

From a theoretical point of view, researchers could focus on novel ways to improve the computational efficiency of estimations of nonlinear model parameters. Any improvement in computational speed can potentially lead to a short-term competitive edge for high-frequency traders who are specialised in volatility-linked exchange traded products. Moreover, to make well-informed decisions based on multi-horizon forecasts, it is necessary to have common measures of predictive ability to evaluate direct or iterated forecasts at multiple horizons jointly rather than considering them individually. Because evaluating multi-horizon forecasting performances of various models individually may potentially lead to spurious results, and thus premature conclusions (Quaedvlieg, 2018).

[14] According to Cboe's website, the VVIX is a volatility of volatility measure in that it represents the expected volatility of the 30-day forward price of the VIX.

§8. Summary and conclusion

This paper evaluates the multi-horizon forecast performances of six GARCH-type models, namely the ARCH, GARCH, GJR-GARCH, CGARCH, APARCH, and GARCHX models, using four equity indices of which two developed markets indices (S&P 500 and HSI) and two emerging markets indices (IPC and KOSPI). Additionally, we use several variance proxies, namely realised volatilities and range-based volatilities, to assess the forecast performances of these GARCH-type models. We evaluate the forecast performances with MSE and QLIKE loss functions. Further, we introduce the Diebold-Mariano test statistic (Diebold & Mariano, 1995) and Model Confidence Set (Hansen et al., 2011) to test the significance of forecast losses at multiple horizons. The aim of this paper is to find out which forecast models generate the most accurate out-of-sample direct forecasts of conditional variances in the short run and in the medium term. In addition, this paper studies the underlying features that drive the performance of GARCH models at multiple horizons. Thus, the research questions are **(1)** *"Which of the following GARCH-type models generate the best volatility forecasts in the short term: ARCH, GARCH, GJR-GARCH, CGARCH, APARCH, or GARCHX?"*, **(2)** *"Which of the following GARCH-type models generate the best volatility forecasts in the medium term: ARCH, GARCH, GJR-GARCH, CGARCH, APARCH, or GARCHX?"*, and **(3)** *"What key features drive the forecast performance of these GARCH-type models at multiple horizons?"*

To answer the research questions properly, we cannot draw general conclusions for all equity indices based on the forecast results due to the variations in variance proxies, forecast horizons, loss functions, and significance tests. Therefore, we need to answer these questions independently for each equity index. For the S&P 500, the undisputed forecast model for both short and longer horizons is the customised GARCHX(1,1,1) model, which incorporates an external regressor based on the VIX Index, irrespective of the loss function. Although the 95% MCS of the S&P 500 includes other forecast models as well for longer horizons under QLIKE, GARCHX is still ranked the highest in terms of MCS p-value. For the Hang Seng Index, depending on the loss function, CGARCH and GARCHX perform well when it comes to short- and medium-term volatility forecasts. Both models are frequently included in $\widehat{\mathcal{M}}_{0.95}^{*}$ under QLIKE. Meanwhile, RMSE-based $\widehat{\mathcal{M}}_{0.95}^{*}$ considers CGARCH as the sole superior model.

In addition, for IPC, we observe mixed results based on the Model Confidence Sets. Under RMSE, the CGARCH is considered as the sole superior model for all horizons in almost all cases, except for the 10-minute subsampled realised proxy and the range proxy based on **Eq. (2)** at $h = 1$, where APARCH is part of $\widehat{\mathcal{M}}_{0.95}^{*}$. However, under QLIKE loss function, a variety of models are selected for the 95%

MCS of IPC at multiple horizons, including CGARCH, APARCH, GJR-GARCH, and GARCHX. Still, the p-value ranks the first two models as highest in almost all cases. Based on the results of forecast losses and depending on the loss function, the component GARCH model tend to generate the lowest forecast errors in most cases. Although in some cases, the APARCH model performs better. For KOSPI, the results of forecast losses favour CGARCH for all horizons under QLIKE, while under RMSE APARCH tends to be the better model for short-term forecasts but is surpassed by CGARCH at longer horizons. Based on the MCS results, APARCH tends to be the sole superior model for short horizons under RMSE, whereas CGARCH seems to be the only model in $\widehat{\mathcal{M}}^*_{0.95}$ for horizons larger than a week. Under QLIKE however, many models are included in $\widehat{\mathcal{M}}^*_{0.95}$ for all horizons.

Ultimately, we can conclude that the empirical results of the conditional variance forecasts of these equity indices are rather mixed, since there is no generic forecast model that shows superior performance in every case for all financial assets. Thus, the accuracy of volatility forecasts are mainly dependent upon factors such as forecast procedure, type of variance estimators, loss functions, significance tests, and even underlying data. Yet, from the results of forecast losses, we can conclude that one model consistently underperforms other GARCH models for all equity indices, namely ARCH(1). This can be explained by the fact that the conditional variance equation of ARCH is simply a linear function of past errors that does not incorporate past conditional variances (Engle, 1982). However, in some cases, the uncomplicated ARCH model is part of $\widehat{\mathcal{M}}^*_{0.95}$, but it does not necessarily mean that it is preferred over other GARCH-type models when it comes to volatility forecasting.

Altogether, we can conclude that several factors affect the forecast performance of GARCH models. For instance, adding an external regressor to the GARCH(1,1) model significantly improves the conditional variance forecasts for the S&P 500, but not necessarily for IPC and KOSPI. Admittedly, we extended the standard GARCH model only with the VIX variance proxy and not with similar volatility indices for the other equity indices, resulting in GARCHX being surpassed by APARCH and CGARCH in terms of forecast performance. In addition, the structure of a conditional variance equation affect the out-of-sample volatility forecasts to some extent. For example, the flexible conditional variance equation of APARCH provides relatively accurate forecasts compared to GARCH(1,1), since APARCH is able to capture empirical regularities related to asset return volatility and nests multiple GARCH-type models (Laurent, 2004). Moreover, the CGARCH decomposes volatility into a long-run component and a short-run component (McMillan & Speight, 2004), allowing the conditional variance to mean revert to a time-varying long-run variance rate (Engle & Lee, 1999). Hence, this could possibly explain the forecast results of APARCH and CGARCH for some emerging markets indices in this paper.

This study adds to existing literature by comparing multi-horizon forecast performances of GARCH-type models based on equity indices from developed markets and emerging markets. Additionally, this paper uses several variance proxies for evaluation of forecast results, such as the 5-minute and 10-minute subsampled realised proxies and two range-based proxies. Furthermore, it examines the underlying features of GARCH models that drive the forecast performance at different horizons.

More importantly, what are the implications of this study? First, finance practitioners, such as quant traders, investment managers, and risk managers, should work with a framework where one conducts volatility forecasts and assesses forecast performances in a holistic manner by using multiple GARCH models, forecast approaches, variance proxies, loss functions, and significance tests. This is necessary due to the variations in forecast performances and likelihood of conflicting results. Thus, this framework can provide a comprehensive overview of forecast results to finance practitioners so that they can make deliberate decisions amidst severe market turmoil. Second, to make the most of mixed forecast results, one could use forecast combinations next to individual forecasts, since combining predictions leads to possible diversification benefits. Compared to individual forecasts, forecast combinations are more robust against misspecification biases and measurement errors in the datasets underlying the individual forecasts (Timmermann, 2006). Third, adding the VIX to a forecast model does not necessarily improve the forecast performance for other equity indices. Therefore, it is essential to include an index-specific volatility measure to the model.

Despite the comprehensive study on GARCH-type models, this paper, unfortunately, has its research limitations as well. For instance, the data sample split is done randomly without any further robustness tests. We also did not use the iterated approach for estimations of conditional variances. Further, the model parameters are all estimated under the normal distribution. Additionally, we evaluated the direct forecasts of GARCH-type models individually at each horizon. This paper does not use specific tests in order to test multistep out-of-sample forecasts jointly at once. Hence, these limitations may have resulted in a one-sided view on the evaluation of forecast performances.

Finally, further research is still relevant in this particular field due to the complexity of this topic. For instance, one could use convoluted nonlinear GARCH models and even extend them with covariates that may improve the models' predictive ability. Additionally, evaluating multivariate GARCH models may provide unique solutions to issues related to multi-asset portfolio management, which is beneficial to institutional investors. Besides, one could focus on novel estimation methods and techniques to improve the computational efficiency of model parameter estimates, or on evaluation tests that allows for joint multi-horizon testing of forecast models at once.

References

Aït-Sahalia, Y., Fan, J., & Li, Y. (2013). The leverage effect puzzle: Disentangling sources of bias at high frequency. *Journal of Financial Economics, Vol. 109, Issue 1, July*, pp. 224-249.

Alizadeh, S., Brandt, M. W., & Diebold, F. X. (2002). Range-Based Estimation of Stochastic Volatility Models. *The Journal of Finance, Vol. 57, No. 3, June*, pp. 1047-1091.

Andersen, T. G., & Bollerslev, T. (1998). Answering the Skeptics: Yes, Standard Volatility Models do Provide Accurate Forecasts. *International Economic Review, Vol. 39, No. 4, November*, pp. 885-905.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F., & Diebold, F. X. (2006). Chapter 15 Volatility and Correlation Forecasting. In G. Elliott, C. W. Granger, & A. Timmermann, *Handbook of Economic Forecasting* (Vol. 1, pp. 777-878). Elsevier.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics, Vol. 61, Issue 1, July*, pp. 43-76.

Baele, L. (2005). Volatility Spillover Effects in European Equity Markets. *Journal of Financial and Quantitative Analysis, Vol. 40, No. 2, June*, pp. 373-401.

Baillie, R. T., Bollerslev, T., & Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, Vol. 74, Issue 1, September*, pp. 3-30.

Bannouh, K., van Dijk, D., & Martens, M. (2009). Range-Based Covariance Estimation Using High-Frequency Data: The Realized Co-Range. *Journal of Financial Econometrics, Vol. 7, No. 4, October*, pp. 341-372.

Bera, A. K., & Higgins, M. L. (1993). ARCH Models: Properties, Estimation and Testing. *Journal of Economic Surveys, Vol. 7, Issue 4, December*, pp. 305-366.

Bernardi, M., & Catania, L. (2018). The Model Confidence Set package for R. *International Journal of Computational Economics and Econometrics, Vol. 8, Issue 2*, pp. 144-158.

Black, F., & Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy, Vol. 81, No. 3, May-June*, pp. 637-654.

Blair, B. J., Poon, S.-H., & Taylor, S. J. (2010). Chapter 88 Forecasting S&P 100 Volatility: The Incremental Information Content of Implied Volatilities and High-Frequency Index Returns. In C.-F. Lee, A. C. Lee, & J. Lee, *Handbook of Quantitative Finance and Risk Management* (pp. 1333-1344). Springer.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, Vol. 31, Issue 3, April*, pp. 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, Vol. 52, Issues 1-2, April-May*, pp. 5-59.

Bollerslev, T., Engle, R. F., & Nelson, D. B. (1994). Chapter 49 ARCH Models. In R. F. Engle, & D. L. McFadden, *Handbook of Econometrics* (Vol. 4, pp. 2959-3038). Elsevier.

Bollerslev, T., Litvinova, J., & Tauchen, G. (2006). Leverage and Volatility Feedback Effects in High-Frequency Data. *Journal of Financial Econometrics, Vol. 4, No. 3, July*, pp. 353-384.

Bollerslev, T., Patton, A. J., & Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics, Vol. 192, Issue 1, May*, pp. 1-18.

Bouchaud, J.-P., Matacz, A., & Potters, M. (2001). Leverage Effect in Financial Markets: The Retarded Volatility Model. *Physical Review Letters, Vol. 87, Issue 22, November*, pp. 228701-228704.

Box, G. E., & Pierce, D. A. (1970). Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *Journal of the American Statistical Association, Vol. 65. No. 332, December*, pp. 1509-1526.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, Vol. 52, No. 3, September*, pp. 345-370.

Brailsford, T. J., & Faff, R. W. (1996). An evaluation of volatility forecasting techniques. *Journal of Banking & Finance, Vol. 20, Issue 3, April*, pp. 419-438.

Brooks, R. D., Faff, R. W., McKenzie, M. D., & Mitchell, H. (2000). A multi-country study of power ARCH models and national stock market returns. *Journal of International Money and Finance, Vol. 19, Issue 3, June*, pp. 377-397.

Brownlees, C. T., Engle, R. F., & Kelly, B. T. (2011). A practical guide to volatility forecasting through calm and storm. *The Journal of Risk, Vol. 14, No. 2, Winter 2011/12*, pp. 3-22.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research, Vol. 33, Issue 2, November*, pp. 261-304.

Catania, L., & Bernardi, M. (2017). Model Confidence Set Procedure. *R package version 0.1.3*, pp. 1-9. Retrieved from https://cran.r-project.org/web/packages/MCS/MCS.pdf

Choudhry, T., & Wu, H. (2008). Forecasting ability of GARCH vs Kalman filter method: evidence from daily UK time-varying beta. *Journal of Forecasting, Vol. 27, Issue 8*, pp. 670-689.

Christensen, K., & Podolskij, M. (2007). Realized range-based estimation of integrated variance. *Journal of Econometrics, Vol. 141, Issue 2, December*, pp. 323-349.

Christiansen, C. (2007). Volatility-Spillover Effects in European Bond Markets. *European Financial Management, Vol. 13, Issue 5, November*, pp. 923-948.

Christie, A. A. (1982). The stochastic behavior of common stock variances: Value, leverage and interest rate effects. *Journal of Financial Economics, Vol. 10, Issue 4, December*, pp. 407-432.

Christoffersen, P. F. (2012). *Elements of Financial Risk Management* (2nd ed.). Academic Press.

Clark, T. E., & McCracken, M. W. (2005). Evaluating Direct Multistep Forecasts. *Econometric Reviews, Vol. 25, Issue 4*, pp. 369-404.

Conrad, C., Karanasos, M., & Zeng, N. (2011). Multivariate fractionally integrated APARCH modeling of stock market volatility: A multi-country study. *Journal of Empirical Finance, Vol. 18, Issue 1, January*, pp. 147-159.

Degiannakis, S. (2004). Volatility forecasting: evidence from a fractional integrated asymmetric power ARCH skewed-t model. *Applied Financial Economics, Vol. 14, Issue 18, December*, pp. 1333-1342.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association, Vol. 74, No. 366, June*, pp. 427-431.

Diebold, F. X. (2015). Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests. *Journal of Business & Economic Statistics, Vol. 33, Issue 1, January*, pp. 1-9.

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics, Vol. 13, No. 3, July*, pp. 253-263.

Diebold, F. X., & Yilmaz, K. (2009). Measuring Financial Asset Return and Volatility Spillovers, with Application to Global Equity Markets. *The Economic Journal, Vol. 119, No. 534, January*, pp. 158-171.

Ding, Z., & Granger, C. W. (1996). Modeling volatility persistence of speculative returns: A new approach. *Journal of Econometrics, Vol. 73, Issue 1, July*, pp. 185-215.

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

Enders, W. (2015). *Applied Econometric Time Series* (4th ed.). John Wiley & Sons.

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica, Vol. 50, No. 4, July*, pp. 987-1007.

Engle, R. F. (1993). Statistical Models for Financial Volatility. *Financial Analysts Journal, Vol. 49, No. 1, January-February*, pp. 72-78.

Engle, R. F., & Lee, G. G. (1999). A Long-Run and Short-Run Component Model of Stock Return Volatility. In R. F. Engle, & H. White, *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W. J. Granger* (pp. 475-501). Oxford University Press.

Engle, R. F., & Patton, A. J. (2001). What good is a volatility model? *Quantitative Finance, Vol. 1, Issue 2*, pp. 237-245.

Fama, E. F. (1965). The Behavior of Stock-Market Prices. *The Journal of Business, Vol. 38, Issue 1, January*, pp. 34-105.

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

Fleming, J., Ostdiek, B., & Whaley, R. E. (1995). Predicting stock market volatility: A new measure. *Journal of Futures Markets, Vol. 15, Issue 3*, pp. 265-302.

Franses, P. H., & Ghijsels, H. (1999). Additive outliers, GARCH and forecasting volatility. *International Journal of Forecasting, Vol. 15, Issue 1, February*, pp. 1-9.

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

Ghalanos, A. (2019). rugarch: Univariate GARCH models. *R package version 1.4.0*, pp. 1-108. Retrieved from https://cran.r-project.org/web/packages/rugarch/rugarch.pdf

Ghysels, E., Rubia, A., & Valkanov, R. (2009). Multi-Period Forecasts of Volatility: Direct, Iterated, and Mixed-Data Approaches. *EFA 2009 Bergen Meetings Paper*, pp. 1-22.

Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *The Journal of Finance, Vol. 48, No. 5, December*, pp. 1779-1801.

Hamao, Y., Masulis, R. W., & Ng, V. (1990). Correlations in Price Changes and Volatility across International Stock Markets. *The Review of Financial Studies, Vol. 3, No. 2*, pp. 281-307.

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

Hansen, P. R., & Lunde, A. (2006). Realized Variance and Market Microstructure. *Journal of Business & Economic Statistics, Vol. 24, No. 2, April*, pp. 127-161.

Hansen, P. R., & Timmermann, A. (2012). Choice of Sample Split in Out-of-Sample Forecast Evaluation. *European University Institute Working Paper, ECO 2012/10*, pp. 1-47.

Hansen, P. R., Lunde, A., & Nason, J. M. (2003). Choosing the Best Volatility Models: The Model Confidence Set Approach. *Oxford Bulletin of Economics and Statistics, Vol. 65, Issue s1*, pp. 839-861.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The Model Confidence Set. *Econometrica, Vol. 79, No. 2, March*, pp. 453-497.

Hao, J., & Zhang, J. E. (2013). GARCH Option Pricing Models, the CBOE VIX, and Variance Risk Premium. *Journal of Financial Econometrics, Vol. 11, Issue 3, June*, pp. 556-580.

Harris, R. I. (1992). Testing for unit roots using the augmented Dickey-Fuller test: Some issues relating to the size, power and the lag structure of the test. *Economics Letters, Vol. 38, Issue 4, April*, pp. 381-386.

Harvey, D. I., Leybourne, S. J., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting, Vol. 13, Issue 2, June*, pp. 281-291.

Hol, E. M. (2003). Chapter 6 Forecasting the Variability of Stock Index Returns with Stochastic Volatility Models and Implied Volatility. In *Empirical Studies on Volatility in International Stock Markets. Dynamic Modeling and Econometrics in Economics and Finance* (Vol. 6, pp. 71-97). Springer.

Hull, J. C. (2012). *Options, Futures, And Other Derivatives* (Global 8th ed.). Pearson Education.

Hull, J. C. (2015). *Risk Management and Financial Institutions* (4th ed.). John Wiley & Sons.

Inoue, A., & Kilian, L. (2005). In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use? *Econometric Reviews, Vol. 23, Issue 4*, pp. 371-402.

Jarque, C. M., & Bera, A. K. (1987). A Test for Normality of Observations and Regression Residuals. *International Statistical Review / Revue Internationale de Statistique, Vol. 55, No. 2, August*, pp. 163-172.

Kambouroudis, D. S., & McMillan, D. G. (2016). Does VIX or volume improve GARCH volatility forecasts? *Applied Economics, Vol. 48, Issue 13*, pp. 1210-1228.

Kay, J. (2007, April 27). Unimaginable horror. *Financial Times*. Retrieved May 30, 2017, from https://www.ft.com/content/824ac36c-f134-11db-838b-000b5df10621

Koutmos, G., & Booth, G. G. (1995). Asymmetric volatility transmission in international stock markets. *Journal of International Money and Finance, Vol. 14, Issue 6, December*, pp. 747-762.

Kuha, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance. *Sociological Methods & Research, Vol. 33, No. 2, November*, pp. 188-229.

Laurent, S. (2004). Analytical Derivates of the APARCH Model. *Computational Economics, Vol. 24, Issue 1, August*, pp. 51-57.

Lintner, J. (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets. *The Review of Economics and Statistics, Vol. 47, No. 1, February*, pp. 13-37.

Ljung, G. M., & Box, G. E. (1978). On a Measure of Lack of Fit in Time Series Models. *Biometrika, Vol. 65, No. 2, August*, pp. 297-303.

Lopez, J. A. (2001). Evaluating the predictive accuracy of volatility models. *Journal of Forecasting, Vol. 20, Issue 2, March*, pp. 87-109.

Mandelbrot, B. (1963). The Variation of Certain Speculative Prices. *The Journal of Business, Vol. 39, No. 4, October*, pp. 394-419.

Marcellino, M., Stock, J. H., & Watson, M. W. (2006). A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. *Journal of Econometrics, Vol. 135, Issues 1-2, November-December*, pp. 499-526.

Mariano, R. S. (2002). Chapter 13 Testing Forecast Accuracy. In M. P. Clements, & D. F. Hendry, *A Campanion to Economic Forecasting* (pp. 284-298). Blackwell Publishing.

Martens, M., & van Dijk, D. (2007). Measuring volality with the realized range. *Journal of Econometrics, Vol. 138, Issue 1, May*, pp. 181-207.

McCracken, M. W., & McGillicuddy, J. (2017). An Empirical Investigation of Direct and Iterated Multistep Conditional Forecasts. *Working Paper 2017-040A, November, Federal Reserve Bank of St. Louis*, pp. 1-40.

McMillan, D. G., & Speight, A. E. (2004). Daily volatility forecasts: reassessing the performance of GARCH models. *Journal of Forecasting, Vol. 23, Issue 6, September*, pp. 449-460.

Merton, R. C. (1973). Theory of Rational Option Pricing. *The Bell Journal of Economics and Management Science, Vol. 4, No. 1, Spring*, pp. 141-183.

Mossin, J. (1966). Equilibrium in a Capital Asset Market. *Econometrica, Vol. 34, No. 4, October*, pp. 768-783.

MSCI. (2018). MSCI Global Investable Market Indexes Methodology - Index Construction Objectives, Guiding Principles and Methodology for the MSCI Global Investable Market Indexes. *Index Methodology, November*, pp. 1-191.

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica, Vol. 59, No. 2, March*, pp. 347-370.

Newey, W. K., & West, K. D. (1994). Automatic Lag Selection in Covariance Matrix Estimation. *The Review of Economic Studies, Vol. 61, Issue 4, October*, pp. 631-653.

Ng, A. (2000). Volatility spillover effects from Japan and the US to the Pacific-Basin. *Journal of International Money and Finance, Vol. 19, Issue 2, April*, pp. 207-233.

Patton, A. J. (2011a). Data-based ranking of realised volatility estimators. *Journal of Econometrics, Vol. 161, Issue 2, April*, pp. 284-303.

Patton, A. J. (2011b). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics, Vol. 160, Issue 1, January*, pp. 246-256.

Poon, S.-H., & Granger, C. W. (2003). Forecasting Volatility in Financial Markets: A Review. *Journal of Economic Literature, Vol. 41, No. 2, June*, pp. 478-539.

Poon, S.-H., & Granger, C. W. (2005). Practical Issues in Forecasting Volatility. *Financial Analysts Journal, Vol. 61, No. 1, January-February*, pp. 45-56.

Poterba, J. M., & Summers, L. H. (1986). The Persistence of Volatility and Stock Market Fluctuations. *The American Economic Review, Vol. 76, No. 5, December*, pp. 1142-1151.

Proietti, T. (2011). Direct and iterated multistep AR methods for difference stationary processes. *International Journal of Forecasting, Vol. 27, Issue 2, April-June*, pp. 266-280.

Quaedvlieg, R. (2018). Multi-Horizon Forecast Comparison. *Working Paper, Erasmus University Rotterdam*, pp. 1-42.

Ross, S. A. (1976). The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory, Vol. 13, Issue 3, December*, pp. 341-360.

Samuels, J. D., & Sekkel, R. M. (2013). Forecasting with Many Models: Model Confidence Sets and Forecast Combination. *Bank of Canada, Working Paper 2013-11*, pp. 1-50.

Schwert, G. W. (1989). Why Does Stock Market Volatility Change Over Time? *Journal of Finance, Vol. 44, Issue 5*, pp. 1115-1153.

Sharpe, W. F. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance, Vol. 19, No. 3, September*, pp. 425-442.

Skintzi, V. D., & Refenes, A. N. (2006). Volatility spillovers and dynamic correlation in European bond markets. *Journal of International Financial Markets, Institutions and Money, Vol. 16, Issue 1, February*, pp. 23-40.

Susmel, R., & Engle, R. F. (1994). Hourly volatility spillovers between international equity markets. *Journal of International Money and Finance, Vol. 13, Issue 1, February*, pp. 3-25.

Timmermann, A. (2006). Chapter 4 Forecast Combinations. In G. Elliott, C. W. Granger, & A. Timmermann, *Handbook of Economic Forecasting* (Vol. 1, pp. 135-196). Elsevier.

Tsay, R. S. (2005). *Analysis of Financial Time Series* (2nd ed.). John Wiley & Sons.

Tse, Y. (1999). Price Discovery and Volatility Spillovers in the DJIA Index and Futures Markets. *The Journal of Futures Markets, Vol. 19, Issue 8, December*, pp. 911-930.

Tse, Y. K. (1998). The Conditional Heteroscedasticity of the Yen-Dollar Exchange Rate. *Journal of Applied Econometrics, Vol. 13, No. 1, January-February*, pp. 49-55.

Watanabe, T., & Harada, K. (2006). Effects of the Bank of Japan's intervention on yen/dollar exchange rate volatility. *Journal of the Japanese and International Economies, Vol. 20, Issue 1, March*, pp. 99-111.

Yarovaya, L., Brzeszczyński, J., & Lau, M. C. (2017). Asymmetry in spillover effects: Evidence for international stock index futures markets. *International Review of Financial Analysis, Vol. 53, October*, pp. 94-111.

Appendix

Section A – Significance tests

• *A.1. Jarque-Bera normality test*

To find out whether a statistical dataset is normally distributed, one can conduct tests for normality of observations such as the Jarque-Bera test (Jarque & Bera, 1987). The null hypothesis, H_0, of this test states that the sample data is normally distributed, i.e. a joint hypothesis in which skewness and kurtosis coefficients are equal to zero and three, respectively. Meanwhile, the alternative hypothesis, H_a, states that the sample data is not normally distributed. As described in Jarque and Bera (1987), the Lagrange multiplier (LM) test statistic, or Jarque-Bera test statistic, is defined by

$$LM = N\left(\frac{\left(\sqrt{b_1}\right)^2}{6} + \frac{(b_2 - 3)^2}{24}\right), \tag{A.1}$$

where N is the number of sample observations, and $\sqrt{b_1}$ and b_2 are the sample skewness and kurtosis coefficients, respectively. It turns out that this test statistic can be compared asymptotically with a chi-squared, χ^2, distribution with two degrees of freedom. For large samples, H_0 is rejected if the LM test statistic exceeds a critical value from the $\chi^2_{(2)}$ distribution (Jarque & Bera, 1987). These critical values can be easily obtained from online sources.[15] With two degrees of freedom, the 1% significance level results in a critical value of 9.21.

• *A.2. Augmented Dickey-Fuller test*

An issue that frequently occurs in time series analysis is the occurrence of nonstationary processes in economic time series. To examine whether unit roots are present in economic time series, the augmented Dickey-Fuller (ADF) test can be conducted (Dickey & Fuller, 1979). The ADF test can be described in two steps. First, consider a simple first order autoregressive (AR) model

$$x_t = \rho x_{t-1} + \varepsilon_t, \tag{A.2}$$

where x_t stands for value of series in period t, while ρ is a coefficient. A unit root is present in a time series when $\rho = 1$, which implies that the series follows a nonstationary process (Harris, 1992). Stationary time series are usually obtained by simply taking the first order, or even second order, difference of the variable of interest. Hence, the next step of the ADF test is

$$\Delta x_t = \phi x_{t-1} + \varepsilon_t, \tag{A.3}$$

[15] χ^2 distribution table source: "http://www.socr.ucla.edu/Applets.dir/ChiSquareTable.html".

which is the first order difference. In this case, the H_0 of the ADF test states that a unit root is present if $\phi = 0$, which is equivalent to testing $\rho = 1$. The alternative hypothesis is $\phi < 0$, which states that a unit root is not present, i.e. the variable of interest follows a stationary process.

The results of the unit root tests are illustrated in **Table A1**, in which sample data of equity indices are tested on both index price level and return level. The ADF test coefficients from **Panel A** show insignificant results for the two Asian equity indices, whereas the two indices from the Americas are statistically significant at the 10% level. Hence, we fail to reject the null hypothesis for the Asian indices, suggesting that these equity indices tend to follow a unit root process on an index price level. Furthermore, **Panel B** displays the ADF test coefficients of logarithmic returns. All of them are statistically significant at the 1% level, which implies that the logarithmic returns of these stock indices tend to follow a stationary process.

Table A1. – Unit root tests for price levels and log returns

Indices	Coefficient	Std. error	*t*-stat	*p*-value
Panel A: Price level				
S&P 500	-0.002*	0.001	-1.853	0.064
HSI	-0.001	0.002	-0.463	0.643
IPC	-0.003*	0.002	-1.680	0.093
KOSPI	-0.002	0.002	-1.451	0.147
Panel B: Log returns				
S&P 500	-1.092***	0.017	-66.133	0
HSI	-1.059***	0.016	-68.238	0
IPC	-0.929***	0.016	-57.487	0
KOSPI	-1.018***	0.016	-62.674	0

Table A1. – This table illustrates the ADF test coefficients of the stock indices based on price levels and log returns, which are represented in **Panels A** and **B**, respectively. In total, 2,948 observations were used in this test. *** and * denote significance at 1% and 10% level, respectively.

In addition, we perform the augmented Dickey-Fuller test on the volatility proxies of these stock indices as well, which are reported in **Table A2**. From this table, we observe statistically significant ADF coefficients at the 1% significance level in each panel. These results suggest that there is sufficient statistical evidence to reject the null hypothesis for all variance proxies in every case. Hence, the variance proxies tend to follow a stationary process because we do not find statistical evidence of the presence of unit roots in these series.

Table A2. – Unit root tests for volatility proxies

Indices	Coefficient	Std. error	*t*-stat	*p*-value
Panel A: *RV* (5-min)				
S&P 500	-0.154***	0.014	-10.799	0
HSI	-0.380***	0.011	-33.297	0
IPC	-0.586***	0.017	-35.311	0
KOSPI	-0.262***	0.013	-19.919	0
Panel B: *RV* (10-min)				
S&P 500	-0.173***	0.015	-11.715	0
HSI	-0.384***	0.011	-33.562	0
IPC	-0.476***	0.018	-27.175	0
KOSPI	-0.303***	0.014	-21.266	0
Panel C: *RP* (Eq. 2)				
S&P 500	-0.411***	0.014	-29.993	0
HSI	-0.551***	0.010	-55.733	0
IPC	-0.532***	0.014	-37.981	0
KOSPI	-0.332***	0.016	-21.021	0
Panel D: *RP* (Eq. 3)				
S&P 500	-0.350***	0.016	-22.084	0
HSI	-0.715***	0.012	-60.927	0
IPC	-0.492***	0.014	-35.290	0
KOSPI	-0.374***	0.014	-26.338	0

Table A2. – This table illustrates the ADF test coefficients of the 5 and 10-minutes realised variances and range-based variances of the four equity indices. In total, 2,948 observations were used in this test. *** denotes significance at 1% level.

• *A.3. Ljung-Box portmanteau test*

Qualitative tools such as the sample autocorrelation and partial autocorrelation functions can easily assess the presence of serial correlation at individual lags. On the other hand, quantitative tools like the Ljung-Box portmanteau test, or Q-test, can be used for testing serial correlation jointly at multiple lags, i.e. it tests the overall randomness of series based on a number of lags instead of testing randomness at each distinct lag (Ljung & Box, 1978). The null of Ljung-Box Q-test states that the first $\mathcal{k}$ serial correlations of a series are jointly equal to zero, which means that the series is independently distributed. In other words, $H_0: \rho_1 = \rho_2 = \cdots = \rho_{\mathcal{k}} = 0$. Meanwhile, the alternative hypothesis states that the series exhibits autocorrelation. The Q-test statistic is formally defined as

$$\tilde{Q}(\mathcal{k}) = n(n+2)\sum_{m=1}^{\mathcal{k}} \frac{\hat{\rho}_m}{n-m}, \qquad \textbf{(A.4)}$$

where n is the sample size, $\hat{\rho}_m$ is the sample autocorrelation at lag m, and $\mathcal{k}$ is the number of lags being tested. When the sample size is sufficiently large, $\tilde{Q}$ could be distributed approximately as $\chi^2_{(\mathcal{k})}$ with $\mathcal{k}$ degrees of freedom (Box & Pierce, 1970). In this case, the null is rejected with significance level α if the Ljung-Box $\tilde{Q} > \chi^2_{(\mathcal{k}),1-\alpha}$, in which $\chi^2_{(\mathcal{k}),1-\alpha}$ is the $1-\alpha$ quantile of the $\chi^2_{(\mathcal{k})}$ distribution.

In **Table A3,** we observe the Ljung-Box Q-test statistics of the variance estimators. We can conclude that there is sufficient statistical evidence to reject the null hypothesis for all cases, which indicates that these volatility proxy series tend to be serially correlated. In accordance with the autocorrelation and partial autocorrelation plots from **Section 4.1**, the Ljung-Box Q-test confirms the occurrence of serial correlation in these series.

Table A3. – Ljung-Box test statistics

	Lag = 1		Lag = 20		Lag = 40	
Indices	$\tilde{Q}(k=1)$	***p*-value**	$\tilde{Q}(k=20)$	***p*-value**	$\tilde{Q}(k=40)$	***p*-value**
Panel A: *RV* (5-min)						
S&P 500	1,041.86***	0	9,641.91***	0	14,777.99***	0
HSI	1,116.47***	0	8,443.70***	0	12,823.92***	0
IPC	278.00***	0	2,858.52***	0	4,011.57***	0
KOSPI	1,357.33***	0	11,293.16***	0	15,761.83***	0
Panel B: *RV* (10-min)						
S&P 500	1,022.73***	0	9,464.34***	0	14,578.30***	0
HSI	1,049.04***	0	7,580.12***	0	11,435.60***	0
IPC	348.04***	0	3,906.69***	0	5,604.53***	0
KOSPI	1,198.31***	0	10,030.04***	0	14,046.23***	0
Panel C: *RP* (Eq. 2)						
S&P 500	880.63***	0	8,271.80***	0	12,941.21***	0
HSI	629.84***	0	3,221.93***	0	4,749.52***	0
IPC	537.13***	0	4,941.63***	0	7,449.07***	0
KOSPI	1,012.11***	0	5,457.85***	0	7,370.89***	0
Panel D: *RP* (Eq. 3)						
S&P 500	752.20***	0	6,659.52***	0	10,328.59***	0
HSI	261.47***	0	2,473.92***	0	3,828.09***	0
IPC	586.12***	0	4,806.33***	0	7,145.75***	0
KOSPI	1,156.94***	0	5,260.44***	0	6,980.69***	0

Table A3. – This table illustrates the results of Ljung-Box Q-test statistics of the volatility proxies. We use three different lag periods in this test, namely one lag, 20 lags, and 40 lags. The critical values corresponding to one, 20, and 40 degrees of freedom at 1% significance level from a $\chi^2_{(k)}$ distribution are 6.635, 37.566, and 63.691, respectively. Furthermore, *** denotes significance at 1% level.

• *A.4. Engle's ARCH test*

A time series revealing serial correlation in the disturbances, also known as conditional heteroskedasticity, is said to have ARCH effects. This phenomenon causes uncorrelated time series to be serially dependent because of a dynamic conditional variance process. To gauge the significance of ARCH effects in economic time series, we can use a Lagrange multiplier test called the ARCH test (Engle, 1982). The null propounds that the squared errors, ξ_t^2, are not serially dependent, i.e. $H_0: \alpha_0 = \alpha_1 = \cdots = \alpha_n = 0$. In contrast, the alternative hypothesis of this test simply states that ξ_t^2 are serially correlated (Tsay, 2005), given by the following regression model $H_a: \xi_t^2 = \alpha_0 + \sum_{i=1}^{k} \alpha_i \, \xi_{t-i}^2 + u_t$. Engle's ARCH test statistic is the common F-statistic for the regression on squared errors. Under the null, the F-

statistic follows a chi-squared distribution with m degrees of freedom (Tsay, 2005). A substantially large critical value suggest rejection of the null in favour of the alternative hypothesis.

As described in Engle (1982), to obtain the ARCH test statistics, we perform the following steps:

Step I. Regress the log return series on a constant using ordinary least squares (OLS).

Step II. Predict the residuals of the logarithmic return series.

Step III. Regress the predicted residuals on a constant using OLS.

Step IV. Perform the ARCH LM test.

In **Table A4,** we find the test statistics of Engle's ARCH LM tests. We use three different lag periods to check for distant ARCH effects in our time series data. The results below suggest that there is sufficient statistical evidence to reject the null for some cases. Note that the ARCH test statistics for IPC and KOSPI indices are statistically insignificant only for the first lag, whereas these ARCH test statistics tend to be significant after five lag periods. Still, these results suggest that the squared residuals of these equity indices tend to be serially dependent, which implies that forecast models that allow for conditional heteroskedasticity could potentially generate more accurate volatility estimates.

Table A4. – ARCH LM test statistics

	Lag = 1		**Lag = 5**		**Lag = 10**	
Indices	**Test statistic**	***p*-value**	**Test statistic**	***p*-value**	**Test statistic**	***p*-value**
S&P 500	8.305***	0.004	40.076***	0	46.021***	0
HSI	4.126***	0.042	19.303***	0.017	31.568***	0.001
IPC	1.072	0.301	38.834***	0	41.608***	0
KOSPI	1.959	0.162	13.289***	0.021	16.485*	0.087

Table A4. – This table illustrates the in-sample results of the ARCH LM test. In this test, we use three different lag periods, namely one, five, and ten lags. The critical values for one, five, and ten degrees of freedom at 1% significance level from a $\chi^2_{(k)}$ distribution are 6.635, 15.086, and 23.209, respectively. Furthermore, *** and * denote the significance at 1% and 10% levels, respectively.

Section B – Supplementary tables

Table B1. – In-sample estimates of model parameters

Forecast model	ω_0	α_j	φ_i	λ_j	x	y	δ	υ_k
Panel A: S&P 500								
ARCH(1)	0	0.932***						
	(0)	(0.025)						
	[0.998]	[0]						
GARCH(1,1)	0	0.069	0.928***					
	(0)	(0.115)	(0.112)					
	[0.931]	[0.548]	[0]					
GJR-GARCH(1,1)	0	0.012	0.938***	0.096				
	(0)	(0.120)	(0.174)	(0.118)				
	[0.950]	[0.923]	[0]	[0.418]				
CGARCH(1,1)	0	0	0.406		0.996***	0.067		
	(0)	(0.146)	(0.820)		(0.009)	(0.055)		
	[0.728]	[1.000]	[0.621]		[0]	[0.228]		
APARCH(1,1)	0	0.046***	0.950***	1.000***			1.129***	
	(0)	(0.001)	(0.005)	(0)			(0.198)	
	[0.271]	[0]	[0]	[0]			[0]	
GARCHX(1,1,1)	0	0.015	0.393					0.458
	(0)	(0.020)	(0.373)					(0.299)
	[1.000]	[0.464]	[0.292]					[0.126]
Panel B: IPC								
ARCH(1)	0***	0.171*						
	(0)	(0.091)						
	[0]	[0.059]						
GARCH(1,1)	0	0.048	0.945					
	(0)	(0.538)	(0.582)					
	[0.980]	[0.930]	[0.104]					
GJR-GARCH(1,1)	0***	0.019	0.857***	0.181***				
	(0)	(0.014)	(0.021)	(0.062)				
	[0]	[0.170]	[0]	[0.004]				
CGARCH(1,1)	0	0.150	0.668***		1.000***	0.020		
	(0)	(0.339)	(0.084)		(0.004)	(0.042)		
	[0.993]	[0.658]	[0]		[0]	[0.633]		
APARCH(1,1)	0.001	0.102***	0.881***	0.687***			0.908**	
	(0.002)	(0.029)	(0.036)	(0.225)			(0.394)	
	[0.581]	[0]	[0]	[0.002]			[0.021]	
GARCHX(1,1,1)	0	0.048	0.945					0
	(0)	(2.085)	(2.446)					(0.198)
	[0.995]	[0.982]	[0.699]					[1.000]

Continued on next page

Forecast model	$\boldsymbol{\omega_0}$	$\boldsymbol{\alpha_j}$	$\boldsymbol{\varphi_i}$	$\boldsymbol{\lambda_j}$	$\boldsymbol{x}$	$\boldsymbol{y}$	$\boldsymbol{\delta}$	$\boldsymbol{\upsilon_k}$
Panel C: HSI								
ARCH(1)	0***	0.144						
	(0)	(0.117)						
	[0]	[0.215]						
GARCH(1,1)	0	0.056	0.924***					
	(0)	(0.041)	(0.022)					
	[0.702]	[0.170]	[0]					
GJR-GARCH(1,1)	0***	0	0.919***	0.105***				
	(0)	(0.005)	(0.009)	(0.030)				
	[0]	[1.000]	[0]	[0]				
CGARCH(1,1)	0	0.044	0.903***		1.000***	0.016		
	(0)	(0.180)	[0.082]		(0.007)	(0.030)		
	[0.994]	[0.808]	[0]		[0]	[0.606]		
APARCH(1,1)	0***	0.032***	0.921***	1.000***			1.785***	
	(0)	(0.007)	(0.024)	(0.001)			(0.097)	
	[0]	[0]	[0]	[0]			[0]	
GARCHX(1,1,1)	0	0.059	0.912***					0.014
	(0)	(0.083)	(0.162)					(0.046)
	[0.845]	[0.477]	[0]					[0.765]
Panel D: KOSPI								
ARCH(1)	0***	0.039						
	(0)	(0.047)						
	[0]	[0.407]						
GARCH(1,1)	0	0.011***	0.986***					
	(0)	(0.002)	(0.003)					
	[0.854]	[0]	[0]					
GJR-GARCH(1,1)	0***	0	0.945***	0.075***				
	(0)	(0.010)	(0.011)	(0.026)				
	[0]	[1.000]	[0]	[0.003]				
CGARCH(1,1)	0	0.075***	0.822***		1.000***	0.011		
	(0)	(0.121)	(0.031)		(0.001)	(0.040)		
	[0.986]	[0.537]	[0]		[0]	[0.787]		
APARCH(1,1)	0.001	0.074***	0.893***	0.880**			0.975***	
	(0.002)	(0.025)	(0.034)	(0.349)			(0.308)	
	[0.482]	[0.003]	[0]	[0.012]			[0.002]	
GARCHX(1,1,1)	0	0.097***	0.770***					0.290
	(0)	(0.024)	(0.124)					(0.255)
	[0.798]	[0]	[0]					[0.256]

Table B1. – This table illustrates the in-sample parameter estimates of the forecast models. Values in parentheses are robust standard errors while values in brackets report p-values of these parameters. The in-sample period starts from January 5, 2000, to August 6, 2004, resulting in 1,000 observations. *** denotes significance at 1% level, ** denotes significance at 5% level, and * denotes significance at 10% level. Note that the parameters are estimated under Gaussian distribution.

Table B2. – Forecast loss coefficients of HSI

	$h = 1$		$h = 5$		$h = 21$		$h = 63$	
Forecast model	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**
Panel A: *RV* (5-min)								
ARCH(1)	9.043	1.212	9.255	1.244	9.293	1.338	9.305	1.119
GARCH(1,1)	3.745	0.481	3.992	0.506	4.353	0.532	4.810	0.603
GJR-GARCH(1,1)	3.795	0.487	4.078	0.511	4.446	0.542	4.909	0.615
CGARCH(1,1)	3.471***	0.462	3.727***	0.487	4.082***	0.516***	4.535***	0.588
APARCH(1,1)	4.068	0.494	4.335	0.516	4.749	0.546	5.181	0.619
GARCHX(1,1,1)	3.859	0.451***	4.145	0.475***	4.493	0.516	4.951	0.581***
Panel B: *RV* (10-min)								
ARCH(1)	9.078	1.235	9.288	1.131	9.284	1.389	9.311	1.079
GARCH(1,1)	3.781	0.506	4.035	0.532	4.404	0.560	4.860	0.628
GJR-GARCH(1,1)	3.826	0.514	4.119	0.539	4.494	0.569	4.960	0.638
CGARCH(1,1)	3.509***	0.487	3.773***	0.513	4.137***	0.543***	4.587***	0.612
APARCH(1,1)	4.090	0.522	4.369	0.545	4.797	0.574	5.227	0.642
GARCHX(1,1,1)	3.886	0.477***	4.186	0.501***	4.541	0.543	5.000	0.606***
Panel C: *RP* (Eq. 2)								
ARCH(1)	9.268	1.262	9.512	1.387	9.413	1.810	9.560	1.260
GARCH(1,1)	4.206	0.680	4.444	0.704	4.864	0.735	5.332	0.807
GJR-GARCH(1,1)	4.198	0.686	4.516	0.711	4.944	0.745	5.424	0.819
CGARCH(1,1)	3.975***	0.659	4.221***	0.683	4.626***	0.716***	5.084***	0.791
APARCH(1,1)	4.401	0.693	4.736	0.716	5.229	0.749	5.672	0.821
GARCHX(1,1,1)	4.265	0.647***	4.576	0.673***	4.985	0.721	5.458	0.784***
Panel D: *RP* (Eq. 3)								
ARCH(1)	9.248	1.281	9.445	1.368	9.409	1.326	9.548	1.262
GARCH(1,1)	4.240	0.643	4.445	0.671	4.821	0.702	5.229	0.767
GJR-GARCH(1,1)	4.265	0.650	4.521	0.678	4.905	0.713	5.320	0.780
CGARCH(1,1)	3.998***	0.623	4.210***	0.650	4.572***	0.683***	4.974***	0.750
APARCH(1,1)	4.500	0.658	4.760	0.683	5.194	0.716	5.577	0.782
GARCHX(1,1,1)	4.330	0.612***	4.593	0.641***	4.946	0.687	5.358	0.744***

Table B2. – This table displays the empirical loss coefficients of various GARCH-type models of the Hang Seng Index at multiple horizons under different variance proxies. For the one-step, five-steps, 21-steps, and 63-steps ahead direct forecasts, we obtained 3,078, 3,074, 3,058, and 3,016 out-of-sample observations, respectively. Note that RMSE coefficients have been scaled up for convenience sake. *** indicates the preferred forecasting model that yields the lowest loss function coefficient, and thus illustrates the best out-of-sample forecast performance in this sample.

Table B3. – Forecast loss coefficients of IPC

	$h=1$		$h=5$		$h=21$		$h=63$	
Forecast model	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**
Panel A: *RV* (5-min)								
ARCH(1)	16.969	1.278	17.003	1.243	17.094	1.241	17.266	1.248
GARCH(1,1)	2.114	0.437	2.306	0.481	2.622	0.549	2.876	0.642
GJR-GARCH(1,1)	2.048	0.421	2.302	0.466	2.667	0.545	2.919	0.628
CGARCH(1,1)	1.987	0.417***	2.168***	0.466	2.441***	0.535***	2.666***	0.625
APARCH(1,1)	1.979***	0.418	2.232	0.464***	2.617	0.542	2.801	0.606***
GARCHX(1,1,1)	2.797	0.421	3.069	0.492	3.407	0.562	3.635	0.642
Panel B: *RV* (10-min)								
ARCH(1)	16.922	1.252	16.948	1.216	17.049	1.214	17.223	1.226
GARCH(1,1)	2.004	0.404	2.223	0.453	2.596	0.528	2.891	0.633
GJR-GARCH(1,1)	1.906	0.390	2.216	0.441	2.639	0.529	2.930	0.620
CGARCH(1,1)	1.904	0.385***	2.107***	0.439***	2.429***	0.516***	2.687***	0.617
APARCH(1,1)	1.848***	0.388	2.157	0.440	2.601	0.526	2.813	0.598***
GARCHX(1,1,1)	2.649	0.393	2.955	0.463	3.351	0.542	3.617	0.630
Panel C: *RP* (Eq. 2)								
ARCH(1)	16.868	1.235	16.962	1.249	17.016	1.277	17.171	1.242
GARCH(1,1)	2.021	0.445	2.164	0.486	2.561	0.563	2.884	0.678
GJR-GARCH(1,1)	1.938	0.430	2.174	0.475	2.582	0.563	2.940	0.663
CGARCH(1,1)	1.928	0.433	2.068***	0.477	2.412***	0.558***	2.693***	0.671
APARCH(1,1)	1.857***	0.428***	2.113	0.472***	2.558	0.562	2.827	0.643***
GARCHX(1,1,1)	2.602	0.453	2.842	0.509	3.268	0.593	3.568	0.690
Panel D: *RP* (Eq. 3)								
ARCH(1)	16.857	1.238	16.986	1.244	17.012	1.260	17.196	1.250
GARCH(1,1)	1.922	0.443	2.109	0.480	2.468	0.546	2.705	0.643
GJR-GARCH(1,1)	1.875	0.434	2.116	0.472	2.510	0.545	2.769	0.632
CGARCH(1,1)	1.777***	0.427***	1.965***	0.467***	2.284***	0.536***	2.492***	0.634
APARCH(1,1)	1.779	0.430	2.043	0.469	2.459	0.542	2.648	0.612***
GARCHX(1,1,1)	2.611	0.451	2.884	0.502	3.241	0.568	3.451	0.655

Table B3. – This table displays the empirical loss coefficients of various GARCH-type models of the IPC index at multiple horizons under different variance proxies. For the one-step, five-steps, 21-steps, and 63-steps ahead direct forecasts, we obtained 3,078, 3,074, 3,058, and 3,016 out-of-sample observations, respectively. Note that RMSE coefficients have been scaled up for convenience sake. *** indicates the preferred forecasting model that yields the lowest loss function coefficient, and thus illustrates the best out-of-sample forecast performance in this sample.

Table B4. – Forecast loss coefficients of KOSPI

	$h = 1$		$h = 5$		$h = 21$		$h = 63$	
Forecast model	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**
Panel A: *RV* (5-min)								
ARCH(1)	10.137	1.374	10.186	2.024	10.241	1.845	10.438	1.324
GARCH(1,1)	1.967	0.403	2.314	0.431	2.959	0.481	3.321	0.556
GJR-GARCH(1,1)	1.918	0.400	2.352	0.436	3.128	0.499	3.542	0.564
CGARCH(1,1)	1.859	0.383***	2.222	0.413***	2.852***	0.467***	3.218***	0.544***
APARCH(1,1)	1.789***	0.402	2.180***	0.437	2.875	0.501	3.250	0.559
GARCHX(1,1,1)	2.770	0.427	2.925	0.451	3.509	0.500	4.051	0.568
Panel B: *RV* (10-min)								
ARCH(1)	10.116	1.408	10.189	1.838	10.245	1.750	10.429	1.274
GARCH(1,1)	2.069	0.418	2.404	0.448	2.996	0.501	3.335	0.576
GJR-GARCH(1,1)	2.004	0.414	2.467	0.452	3.178	0.519	3.558	0.583
CGARCH(1,1)	1.963	0.399***	2.314	0.432***	2.891***	0.488***	3.233***	0.565***
APARCH(1,1)	1.869***	0.416	2.292***	0.453	2.926	0.521	3.267	0.577
GARCHX(1,1,1)	2.819	0.440	3.007	0.467	3.565	0.518	4.065	0.587
Panel C: *RP* (Eq. 2)								
ARCH(1)	10.200	1.383	10.230	2.096	10.396	1.760	10.600	1.409
GARCH(1,1)	2.610	0.583	2.961	0.614	3.526	0.669	3.813	0.742
GJR-GARCH(1,1)	2.358	0.574	2.963	0.616	3.698	0.689	4.011	0.749
CGARCH(1,1)	2.531	0.564***	2.897	0.597***	3.435***	0.654***	3.723***	0.730***
APARCH(1,1)	2.319***	0.574	2.819***	0.616	3.484	0.692	3.754	0.742
GARCHX(1,1,1)	3.126	0.601	3.341	0.629	4.009	0.689	4.453	0.752
Panel D: *RP* (Eq. 3)								
ARCH(1)	10.305	1.298	10.238	2.372	10.405	1.929	10.621	1.406
GARCH(1,1)	2.605	0.561	2.980	0.594	3.548	0.648	3.809	0.718
GJR-GARCH(1,1)	2.398	0.554	2.960	0.597	3.722	0.670	4.013	0.729
CGARCH(1,1)	2.524	0.541***	2.914	0.576***	3.456***	0.634***	3.719***	0.704***
APARCH(1,1)	2.374***	0.555	2.809***	0.598	3.500	0.673	3.758	0.724
GARCHX(1,1,1)	3.194	0.582	3.370	0.613	4.036	0.668	4.458	0.730

Table B4. – This table displays the empirical loss coefficients of various GARCH-type models of KOSPI at multiple horizons under different variance proxies. For the one-step, five-steps, 21-steps, and 63-steps ahead direct forecasts, we obtained 3,078, 3,074, 3,058, and 3,016 out-of-sample observations, respectively. Note that RMSE coefficients have been scaled up for convenience sake. *** indicates the preferred forecasting model that yields the lowest loss function coefficient, and thus illustrates the best out-of-sample forecast performance in this sample.

Table B5. – Coefficients of determination

	***RV* (5-min)**	***RV* (10-min)**	***RP* (Eq. 2)**	***RP* (Eq. 3)**
Indices	R^2	R^2	R^2	R^2
S&P 500	0.594	0.591	0.548	0.480
HSI	0.429	0.409	0.270	0.245
IPC	0.294	0.339	0.349	0.347
KOSPI	0.478	0.460	0.341	0.316

Table B5. – This matrix illustrates the adjusted R^2 coefficients of the out-of-sample regressions of various volatility proxies on the VIX variance proxy. 3,078 out-of-sample observations were used in these regressions.

Table B6. – Diebold-Mariano test statistics of one-day ahead forecasts of S&P 500

Forecast model	**RMSE**						**QLIKE**					
Panel A: *RV* (5-min)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	5.02***	-					8.46***	-				
GJR-GARCH(1,1)	4.96***	3.62***	-				8.80***	7.37***	-			
CGARCH(1,1)	4.98***	3.39***	-3.72***	-			8.96***	2.38**	-1.51	-		
APARCH(1,1)	4.93***	3.36***	0.90	3.34***	-		9.14***	3.91***	0.36	2.88***	-	
GARCHX(1,1,1)	4.89***	3.53***	3.25***	3.57***	3.75***	-	9.48***	3.49***	1.93*	3.59***	2.14**	-
Panel B: *RV* (10-min)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	4.98***	-					8.90***	-				
GJR-GARCH(1,1)	4.94***	3.65***	-				9.27***	6.19***	-			
CGARCH(1,1)	4.94***	3.34***	-3.78***	-			9.42***	2.32**	-1.37	-		
APARCH(1,1)	4.90***	3.38***	1.18	3.38***	-		9.63***	4.23***	0.37	2.92***	-	
GARCHX(1,1,1)	4.84***	3.46***	3.03***	3.50***	3.49***	-	9.79***	3.62***	1.92*	3.77***	2.16**	-
Panel C: *RP* (Eq. 2)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	4.93***	-					13.14***	-				
GJR-GARCH(1,1)	4.89***	3.55***	-				14.50***	5.90***	-			
CGARCH(1,1)	4.89***	3.36***	-3.59***	-			13.57***	2.27**	-1.71*	-		
APARCH(1,1)	4.85***	3.24***	1.21	3.12***	-		14.38***	3.65***	0.28	2.49**	-	
GARCHX(1,1,1)	4.76***	3.45***	3.34***	3.47***	3.57***	-	12.21***	4.16***	2.42**	4.36***	2.61***	-
Panel D: *RP* (Eq. 3)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	5.01***	-					13.39***	-				
GJR-GARCH(1,1)	4.98***	3.50***	-				14.86***	5.45***	-			
CGARCH(1,1)	4.96***	3.40***	-0.81	-			13.83***	2.70***	-1.04	-		
APARCH(1,1)	4.93***	3.11***	0.94	1.85*	-		14.92***	3.02***	0.10	1.29	-	
GARCHX(1,1,1)	4.80***	3.45***	3.40***	3.47***	3.59***	-	11.72***	3.75***	2.24**	3.66***	2.39**	-

Table B6. – This table illustrates the Diebold-Mariano test results of one-day ahead forecasts of S&P 500 for various GARCH-type models. As explained in Diebold (2015), this test statistic can be calculated through regression of the loss differential on a constant with Newey-West standard errors. A positive DM test statistic implies that the column model has the tendency to generate less accurate forecasts compared to the corresponding row model, whereas a negative DM test statistic implies that the column model tends to predict more accurately than the corresponding row model. Statistical significance of the Diebold-Mariano test statistics are determined by the conventional two-tailed critical values of a standard normal distribution. *, **, and *** denote significance at 10%, 5%, and 1% levels, respectively. For the one-step ahead forecasts, 3,078 out-of-sample observations were used to determine the corresponding DM test statistics.

Table B7. – Diebold-Mariano test statistics of one-week ahead forecasts of S&P 500

Forecast model	**RMSE**						**QLIKE**					
Panel A: *RV* (5-min)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	5.02***	-					9.26***	-				
GJR-GARCH(1,1)	5.03***	3.51***	-				9.48***	2.05**	-			
CGARCH(1,1)	4.97***	3.38***	1.84*	-			9.65***	2.00**	-0.92	-		
APARCH(1,1)	4.97***	3.04***	0.90	-4.77***	-		9.52***	2.25**	0.32	1.51	-	
GARCHX(1,1,1)	4.88***	3.39***	3.17***	3.39***	3.52***	-	9.93***	2.62***	1.49	2.44**	2.05**	-
Panel B: *RV* (10-min)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	4.98***	-					9.41***	-				
GJR-GARCH(1,1)	4.99***	3.57***	-				9.65***	2.67***	-			
CGARCH(1,1)	4.94***	3.36***	1.63	-			9.79***	2.04**	-1.18	-		
APARCH(1,1)	4.94***	3.09***	0.89	-4.09***	-		9.59***	2.24**	0.12	1.54	-	
GARCHX(1,1,1)	4.85***	3.37***	3.09***	3.36***	3.47***	-	9.92***	2.69**	1.51	2.56**	2.18**	-
Panel C: *RP* (Eq. 2)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	5.00***	-					10.83***	-				
GJR-GARCH(1,1)	5.01***	3.35***	-				11.63***	3.13***	-			
CGARCH(1,1)	4.95***	3.38***	2.81***	-			11.42***	2.51**	-1.36	-		
APARCH(1,1)	4.96***	2.92***	1.34	-4.38***	-		11.42***	2.44**	-0.07	1.37	-	
GARCHX(1,1,1)	4.83***	3.37***	3.26***	3.36***	3.50***	-	10.98***	2.44**	1.41	2.23**	1.83*	-
Panel D: *RP* (Eq. 3)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	5.05***	-					12.18***	-				
GJR-GARCH(1,1)	5.06***	3.12***	-				13.72***	3.31***	-			
CGARCH(1,1)	4.99***	3.39***	3.10***	-			12.76***	2.96***	-0.94	-		
APARCH(1,1)	5.01***	2.66***	0.87	-4.15***	-		12.74***	1.95*	-0.34	0.45	-	
GARCHX(1,1,1)	4.83***	3.40***	3.35***	3.40***	3.52***	-	10.08***	2.35**	1.39	2.01**	1.84*	-

Table B7. – This table illustrates the Diebold-Mariano test results of one-week ahead forecasts of S&P 500 for various GARCH-type models. As explained in Diebold (2015), this test statistic can be calculated through regression of the loss differential on a constant with Newey-West standard errors. A positive DM test statistic implies that the column model has the tendency to generate less accurate forecasts compared to the corresponding row model, whereas a negative DM test statistic implies that the column model tends to predict more accurately than the corresponding row model. Statistical significance of the Diebold-Mariano test statistics are determined by the conventional two-tailed critical values of a standard normal distribution. *, **, and *** denote significance at 10%, 5%, and 1% levels, respectively. For the one-week ahead forecasts, 3,074 out-of-sample observations were used to determine the corresponding DM test statistics.

Table B8. – Diebold-Mariano test statistics of one-month ahead forecasts of S&P 500

Forecast model	**RMSE**						**QLIKE**					
Panel A: *RV* (5-min)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	5.16***	-					5.52***	-				
GJR-GARCH(1,1)	5.18***	-2.37**	-				5.22***	-2.43**	-			
CGARCH(1,1)	5.09***	3.33***	3.43***	-			4.93***	-2.54**	-1.23	-		
APARCH(1,1)	5.14***	-4.89***	-0.18	-3.77***	-		4.17***	-1.58	-1.36	-1.04	-	
GARCHX(1,1,1)	4.92***	3.32***	3.36***	3.32***	3.46***	-	6.46***	1.68*	2.81***	2.56**	2.37**	-
Panel B: *RV* (10-min)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	5.13***	-					5.67***	-				
GJR-GARCH(1,1)	5.15***	-2.70***	-				5.35***	-2.65***	-			
CGARCH(1,1)	5.06***	3.32***	3.40***	-			5.10***	-2.65***	-1.10	-		
APARCH(1,1)	5.11***	-5.95***	-0.18	-3.72***	-		4.26***	-1.70*	-1.43	-1.16	-	
GARCHX(1,1,1)	4.89***	3.30***	3.33***	3.29***	3.44***	-	6.71***	1.70*	2.81***	2.53**	2.44**	-
Panel C: *RP* (Eq. 2)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	5.10***	-					5.12***	-				
GJR-GARCH(1,1)	5.13***	-2.98***	-				4.89***	-1.88*	-			
CGARCH(1,1)	5.03***	3.32***	3.40***	-			4.77***	-2.40**	-1.36	-		
APARCH(1,1)	5.08***	-5.95***	-0.12	-3.76***	-		4.00***	-1.54	-1.43	-1.14	-	
GARCHX(1,1,1)	4.85***	3.32***	3.35***	3.32***	3.46***	-	5.87***	2.31**	3.20***	3.29***	2.52**	-
Panel D: *RP* (Eq. 3)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	5.15***	-					8.83***	-				
GJR-GARCH(1,1)	5.17***	-3.65***	-				8.47***	-1.54	-			
CGARCH(1,1)	5.07***	3.35***	3.50***	-			7.60***	-2.05**	-1.35	-		
APARCH(1,1)	5.12***	-5.96***	-0.19	-3.82***	-		6.14***	-1.58	-1.54	-1.13	-	
GARCHX(1,1,1)	4.86***	3.34***	3.39***	3.34***	3.48***	-	9.69***	2.89***	3.57***	3.55***	2.87***	-

Table B8. – This table illustrates the Diebold-Mariano test results of one-month ahead forecasts of S&P 500 for various GARCH-type models. As explained in Diebold (2015), this test statistic can be calculated through regression of the loss differential on a constant with Newey-West standard errors. A positive DM test statistic implies that the column model has the tendency to generate less accurate forecasts compared to the corresponding row model, whereas a negative DM test statistic implies that the column model tends to predict more accurately than the corresponding row model. Statistical significance of the Diebold-Mariano test statistics are determined by the conventional two-tailed critical values of a standard normal distribution. *, **, and *** denote significance at 10%, 5%, and 1% levels, respectively. For the one-month ahead forecasts, 3,058 out-of-sample observations were used to determine the corresponding DM test statistics.

Table B9. – Diebold-Mariano test statistics of one-quarter ahead forecasts of S&P 500

Forecast model	**RMSE**						**QLIKE**					
Panel A: *RV* (5-min)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	5.23***	-					3.79***	-				
GJR-GARCH(1,1)	5.27***	-3.60***	-				3.54***	-0.78	-			
CGARCH(1,1)	5.15***	3.35***	3.44***	-			3.35***	-0.95	-0.25	-		
APARCH(1,1)	5.22***	-4.39***	0.79	-3.61***	-		3.43***	-0.57	-0.08	0.17	-	
GARCHX(1,1,1)	4.89***	3.29***	3.32***	3.27***	3.38***	-	4.39***	1.40	1.32	1.26	1.12	-
Panel B: *RV* (10-min)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	5.19***	-					4.06***	-				
GJR-GARCH(1,1)	5.23***	-3.56***	-				3.78***	-0.67	-			
CGARCH(1,1)	5.11***	3.34***	3.42***	-			3.58***	-0.88	-0.25	-		
APARCH(1,1)	5.18***	-4.31***	0.79	-3.59***	-		3.65***	-0.49	-0.05	0.19	-	
GARCHX(1,1,1)	4.86***	3.29***	3.32***	3.27***	3.37***	-	4.71***	1.47	1.26	1.26	1.06	-
Panel C: *RP* (Eq. 2)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	5.20***	-					5.08***	-				
GJR-GARCH(1,1)	5.25***	-3.63***	-				4.85***	-0.48	-			
CGARCH(1,1)	5.12***	3.33***	3.44***	-			4.57***	-0.61	-0.14	-		
APARCH(1,1)	5.19***	-4.40***	0.67	-3.62***	-		4.64***	-0.60	-0.61	-0.34	-	
GARCHX(1,1,1)	4.85***	3.29***	3.32***	3.27***	3.38***	-	6.03***	1.56	1.33	1.30	1.33	-
Panel D: *RP* (Eq. 3)	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**	**ARCH**	**GARCH**	**GJR-GARCH**	**CGARCH**	**APARCH**	**GARCHX**
ARCH(1)	-						-					
GARCH(1,1)	5.21***	-					6.09***	-				
GJR-GARCH(1,1)	5.25***	-3.77***	-				6.04***	-0.09	-			
CGARCH(1,1)	5.12***	3.34***	3.48***	-			5.63***	-0.19	-0.09	-		
APARCH(1,1)	5.19***	-4.74***	0.68	-3.67***	-		5.82***	-0.36	-0.68	-0.35	-	
GARCHX(1,1,1)	4.82***	3.30***	3.34***	3.28***	3.39***	-	7.02***	2.08**	1.52	1.55	1.52	-

Table B9. – This table illustrates the Diebold-Mariano test results of one-quarter ahead forecasts of S&P 500 for various GARCH-type models. As explained in Diebold (2015), this test statistic can be calculated through regression of the loss differential on a constant with Newey-West standard errors. A positive DM test statistic implies that the column model has the tendency to generate less accurate forecasts compared to the corresponding row model, whereas a negative DM test statistic implies that the column model tends to predict more accurately than the corresponding row model. Statistical significance of the Diebold-Mariano test statistics are determined by the conventional two-tailed critical values of a standard normal distribution. *, **, and *** denote significance at 10%, 5%, and 1% levels, respectively. For the one-quarter ahead forecasts, 3,016 out-of-sample observations were used to determine the corresponding DM test statistics.

Table B10. – Results of MCS for HSI

	$h = 1$		$h = 5$		$h = 21$		$h = 63$	
Forecast model	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**
Panel A: *RV* (5-min)								
ARCH(1)	0	0	0	0	0	0	0	0
GARCH(1,1)	0	0	0	0	0	0	0	0
GJR-GARCH(1,1)	0	0	0	0	0	0	0	0
CGARCH(1,1)	1.000***	0.381***	1.000***	0.487***	1.000***	1.000***	1.000***	0.744***
APARCH(1,1)	0	0	0	0	0	0	0	0
GARCHX(1,1,1)	0	1.000***	0	1.000***	0	1.000***	0	1.000***
Panel B: *RV* (10-min)								
ARCH(1)	0	0	0	0	0	0.265***	0	0
GARCH(1,1)	0	0	0	0	0	0	0	0
GJR-GARCH(1,1)	0	0	0	0	0	0.016	0	0
CGARCH(1,1)	1.000***	0.480***	1.000***	0.506***	1.000***	1.000***	1.000***	0.738***
APARCH(1,1)	0	0	0	0	0	0.044	0	0
GARCHX(1,1,1)	0	1.000***	0	1.000***	0	1.000***	0	1.000***
Panel C: *RP* (Eq. 2)								
ARCH(1)	0	0	0	0	0	0.419***	0	0
GARCH(1,1)	0	0	0	0	0	0	0	0
GJR-GARCH(1,1)	0	0	0	0	0	0.010	0	0
CGARCH(1,1)	1.000***	0.458***	1.000***	0.580***	1.000***	1.000***	1.000***	0.679***
APARCH(1,1)	0	0	0	0	0	0.035	0	0
GARCHX(1,1,1)	0	1.000***	0	1.000***	0	1.000***	0	1.000***
Panel D: *RP* (Eq. 3)								
ARCH(1)	0	0	0	0	0	0	0	0
GARCH(1,1)	0	0	0	0	0	0	0	0
GJR-GARCH(1,1)	0	0	0	0	0	0	0	0
CGARCH(1,1)	1.000***	0.398***	1.000***	0.592***	1.000***	1.000***	1.000***	0.723***
APARCH(1,1)	0	0	0	0	0	0	0	0
GARCHX(1,1,1)	0	1.000***	0	1.000***	0	0.968***	0	1.000***

Table B10. – This table provides the p-values of Model Confidence Sets for the Hang Seng Index forecast losses at multiple horizons. A block bootstrap approach of 1,000 resamples, block length equal to forecast horizon, and 5% significance level are used to determine these coefficients. Note that we set $\mathcal{L} = 2$ for the one-step ahead forecasts. *** denotes the forecast model is part of $\widehat{\mathcal{M}}^*_{0.95}$.

Table B11. – Results of MCS for IPC

	$h = 1$		$h = 5$		$h = 21$		$h = 63$	
Forecast model	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**
Panel A: *RV* (5-min)								
ARCH(1)	0	0	0	0	0	0	0	0
GARCH(1,1)	0	0	0	0.002	0	0.024	0	0.147***
GJR-GARCH(1,1)	0	0.729***	0	0.942***	0	0.804***	0	0.099***
CGARCH(1,1)	1.000***	1.000***	1.000***	0.999***	1.000***	1.000***	1.000***	0.811***
APARCH(1,1)	0	0.999***	0	1.000***	0	0.978***	0	1.000***
GARCHX(1,1,1)	0	0.986***	0	0.216***	0	0.416***	0	0.497***
Panel B: *RV* (10-min)								
ARCH(1)	0	0	0	0	0	0	0	0
GARCH(1,1)	0	0	0	0.001	0	0.024	0	0.302***
GJR-GARCH(1,1)	0	0.918***	0	0.996***	0	0.701***	0	0.211***
CGARCH(1,1)	0	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***	0.864***
APARCH(1,1)	1.000***	0.982***	0	1.000***	0	0.955***	0	1.000***
GARCHX(1,1,1)	0	0.836***	0	0.141***	0	0.487***	0	0.519***
Panel C: *RP* (Eq. 2)								
ARCH(1)	0	0	0	0	0	0	0	0
GARCH(1,1)	0	0.008	0	0.082***	0	0.638***	0	0.600***
GJR-GARCH(1,1)	0	0.944***	0	0.978***	0	0.968***	0	0.251***
CGARCH(1,1)	0	0.930***	1.000***	0.975***	1.000***	1.000***	1.000***	0.613***
APARCH(1,1)	1.000***	1.000***	0	1.000***	0	1.000***	0	1.000***
GARCHX(1,1,1)	0	0.296***	0	0.067***	0	0.205***	0	0.155***
Panel D: *RP* (Eq. 3)								
ARCH(1)	0	0	0	0	0	0	0	0
GARCH(1,1)	0	0.002	0	0.031	0	0.038	0	0.617***
GJR-GARCH(1,1)	0	0.917***	0	0.969***	0	0.844***	0	0.227***
CGARCH(1,1)	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***	0.757***
APARCH(1,1)	0	0.993***	0	1.000***	0	0.986***	0	1.000***
GARCHX(1,1,1)	0	0.234***	0	0.065***	0	0.168***	0	0.221***

Table B11. – This table provides the p-values of Model Confidence Sets for the IPC forecast losses at multiple horizons. A block bootstrap approach of 1,000 resamples, block length equal to forecast horizon, and 5% significance level are used to determine these coefficients. Note that we set $\mathcal{L} = 2$ for the one-step ahead forecasts. *** denotes the forecast model is part of $\widehat{\mathcal{M}}^*_{0.95}$.

Table B12. – Results of MCS for KOSPI

	$h = 1$		$h = 5$		$h = 21$		$h = 63$	
Forecast model	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**	**RMSE**	**QLIKE**
Panel A: *RV* (5-min)								
ARCH(1)	0	0	0	0.369***	0	0.275***	0	0
GARCH(1,1)	0	0	0	0	0	0.004	0	0.045
GJR-GARCH(1,1)	0	0.209***	0	0.013	0	0.045	0	0.377***
CGARCH(1,1)	0	1.000***	0.750***	1.000***	1.000***	1.000***	1.000***	1.000***
APARCH(1,1)	1.000***	0.306***	1.000***	0.050***	0	0.077***	0	0.776***
GARCHX(1,1,1)	0	0.209***	0	0.369***	0	1.000***	0	1.000***
Panel B: *RV* (10-min)								
ARCH(1)	0	0	0	0.316***	0	0.280***	0	0
GARCH(1,1)	0	0	0	0	0	0.003	0	0.205***
GJR-GARCH(1,1)	0	0.333***	0	0.085***	0	0.057***	0	0.451***
CGARCH(1,1)	0	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***	1.000***
APARCH(1,1)	1.000***	0.385***	0.235***	0.135***	0	0.098***	0	0.862***
GARCHX(1,1,1)	0	0.333***	0	0.316***	0	1.000***	0	1.000***
Panel C: *RP* (Eq. 2)								
ARCH(1)	0	0	0	0.379***	0	0	0	0
GARCH(1,1)	0	0	0	0	0	0.001	0	0.069***
GJR-GARCH(1,1)	0	0.606***	0	0.213***	0	0.057***	0	0.420***
CGARCH(1,1)	0	1.000***	0	1.000***	1.000***	1.000***	1.000***	1.000***
APARCH(1,1)	1.000***	0.723***	1.000***	0.415***	0	0.092***	0	0.893***
GARCHX(1,1,1)	0	0.606***	0	1.000***	0	1.000***	0	1.000***
Panel D: *RP* (Eq. 3)								
ARCH(1)	0	0	0	0.493***	0	0.274***	0	0
GARCH(1,1)	0	0	0	0	0	0.003	0	0.044
GJR-GARCH(1,1)	0	0.404***	0	0.048	0	0.053***	0	0.165***
CGARCH(1,1)	0	1.000***	0	1.000***	1.000***	1.000***	1.000***	1.000***
APARCH(1,1)	1.000***	0.478***	1.000***	0.140***	0	0.096***	0	0.575***
GARCHX(1,1,1)	0	0.404***	0	0.493***	0	1.000***	0	1.000***

Table B12. – This table provides the p-values of Model Confidence Sets for the KOSPI forecast losses at multiple horizons. A block bootstrap approach of 1,000 resamples, block length equal to forecast horizon, and 5% significance level are used to determine these coefficients. Note that we set $\mathcal{L} = 2$ for the one-step ahead forecasts. *** denotes the forecast model is part of $\widehat{\mathcal{M}}^*_{0.95}$.

Section C – Supplementary graphs

Figure C1. – Annualised volatility proxies of HSI index

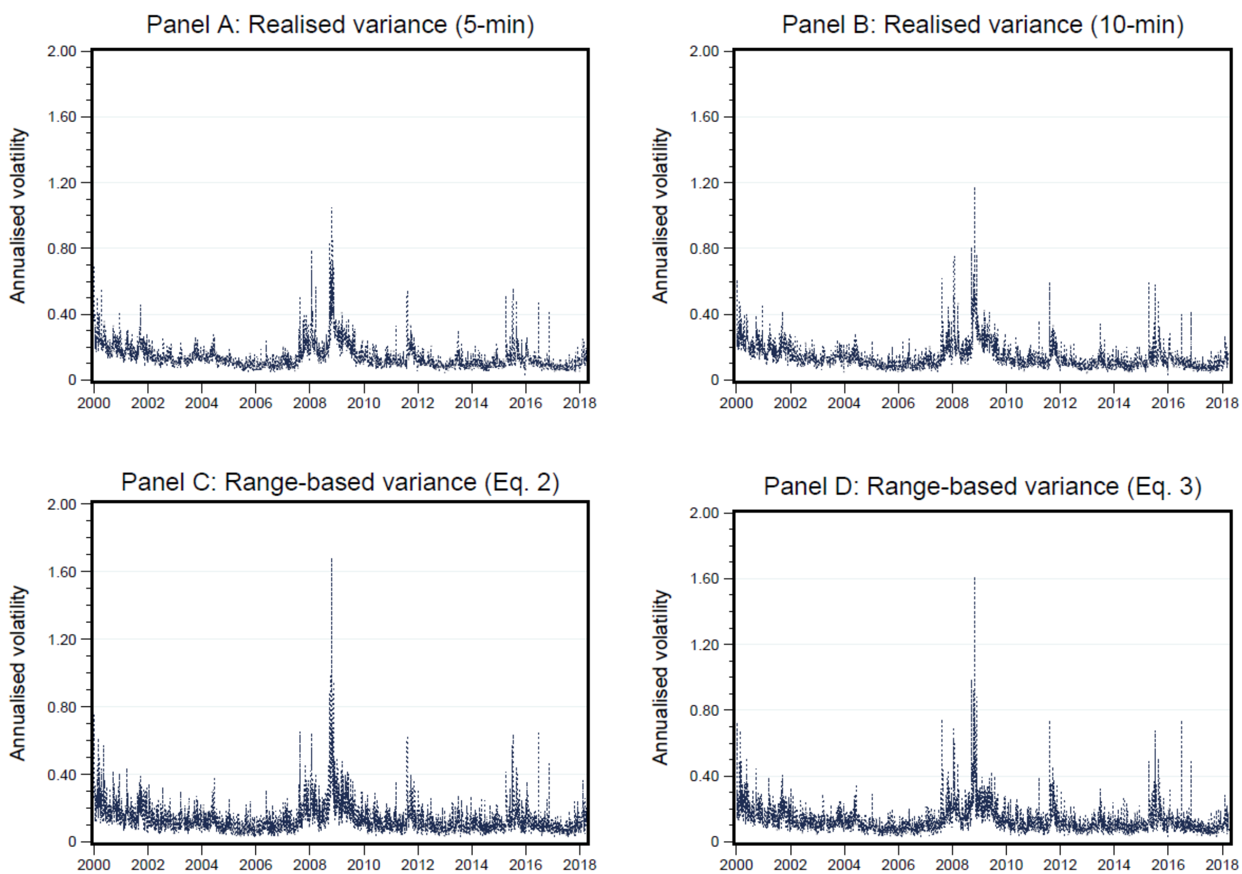

Figure C1. – This figure displays volatility estimates of the HSI index based on four different proxies, namely the 5-minute and 10-minute *RV*, and the two range-based variances based on **Eq. (2)** and **Eq. (3)**. The volatility estimates are annualised, using the following formula from Patton (2011a): $\sigma_t^A = \sqrt{252 \times \widetilde{VAR}_t}$. Furthermore, the vertical axes in all panels are scaled up by 1/100. The sample period starts from January 5, 2000, to March 27, 2018.

Figure C2. – Annualised volatility proxies of IPC index

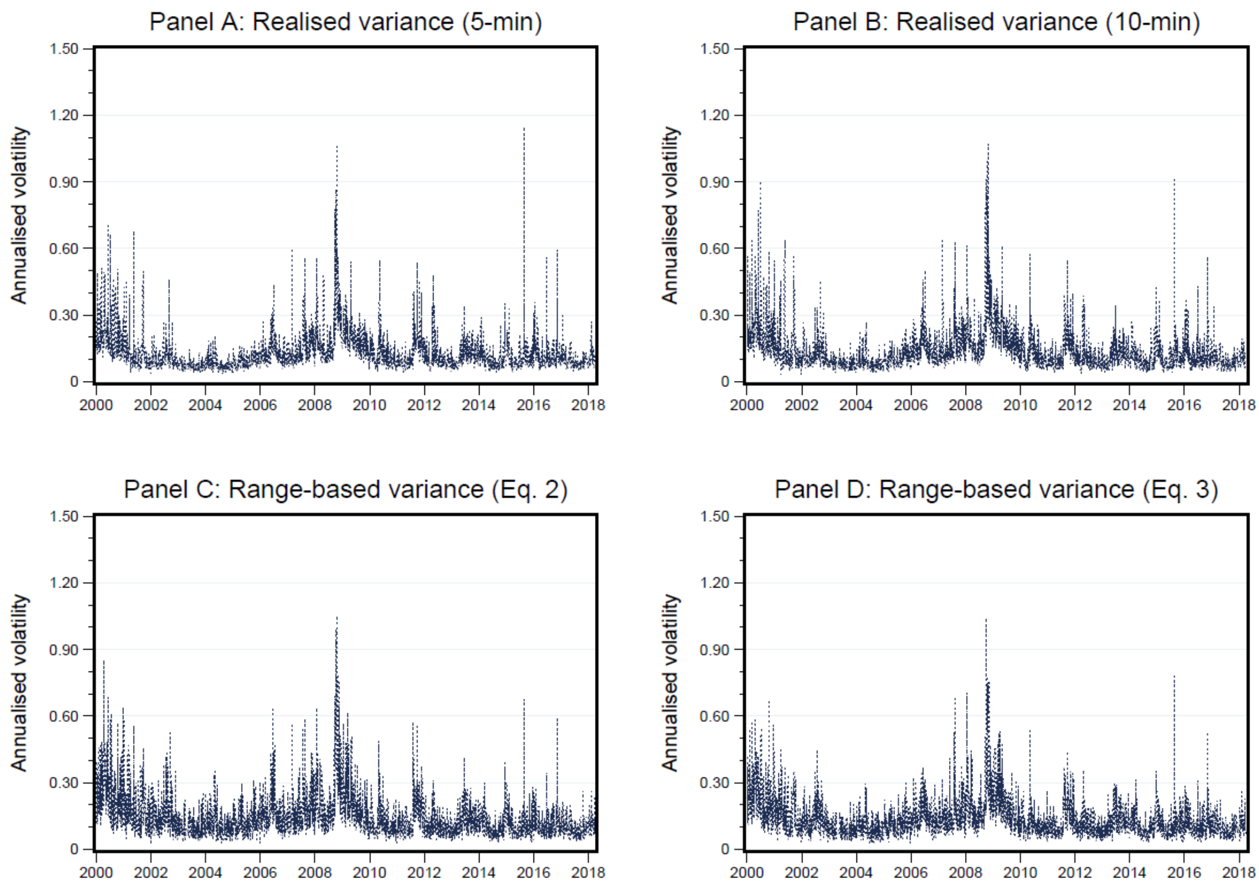

Figure C2. – This figure displays volatility estimates of the IPC index based on four different proxies, namely the 5-minute and 10-minute *RV*, and the two range-based variances based on **Eq. (2)** and **Eq. (3)**. The volatility estimates are annualised, using the following formula from Patton (2011a): $\sigma_t^A = \sqrt{252 \times \widehat{VAR}_t}$. Furthermore, the vertical axes in all panels are scaled up by 1/100. The sample period starts from January 5, 2000, to March 27, 2018.

Figure C3. – Annualised volatility proxies of KOSPI index

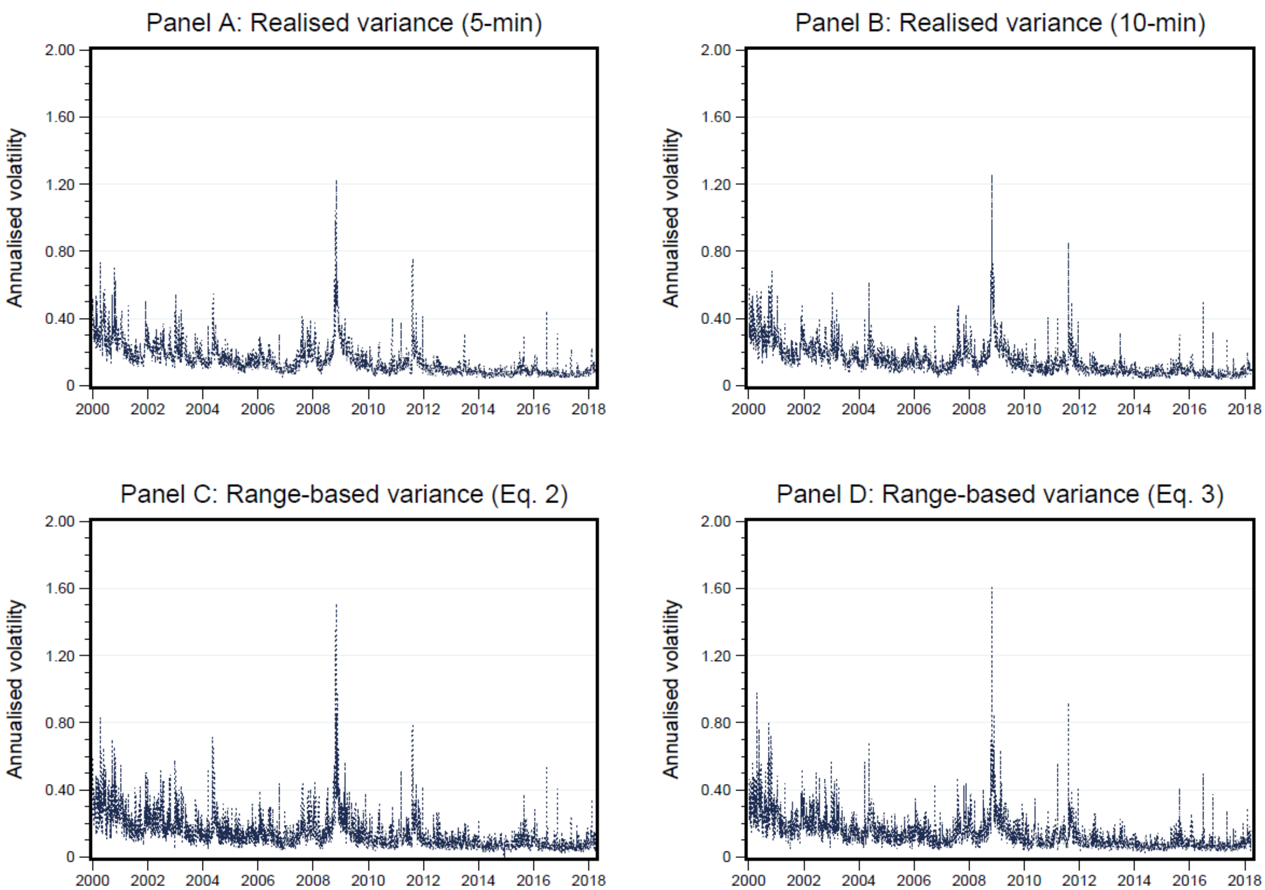

Figure C3. – This figure displays volatility estimates of the KOSPI index based on four different proxies, namely the 5-minute and 10-minute *RV*, and the two range-based variances based on **Eq. (2)** and **Eq. (3)**. The volatility estimates are annualised, using the following formula from Patton (2011a): $\sigma_t^A = \sqrt{252 \times \widehat{VAR}_t}$. Furthermore, the vertical axes in all panels are scaled up by 1/100. The sample period starts from January 5, 2000, to March 27, 2018.

Figure C4. – Autocorrelations of HSI volatility proxies

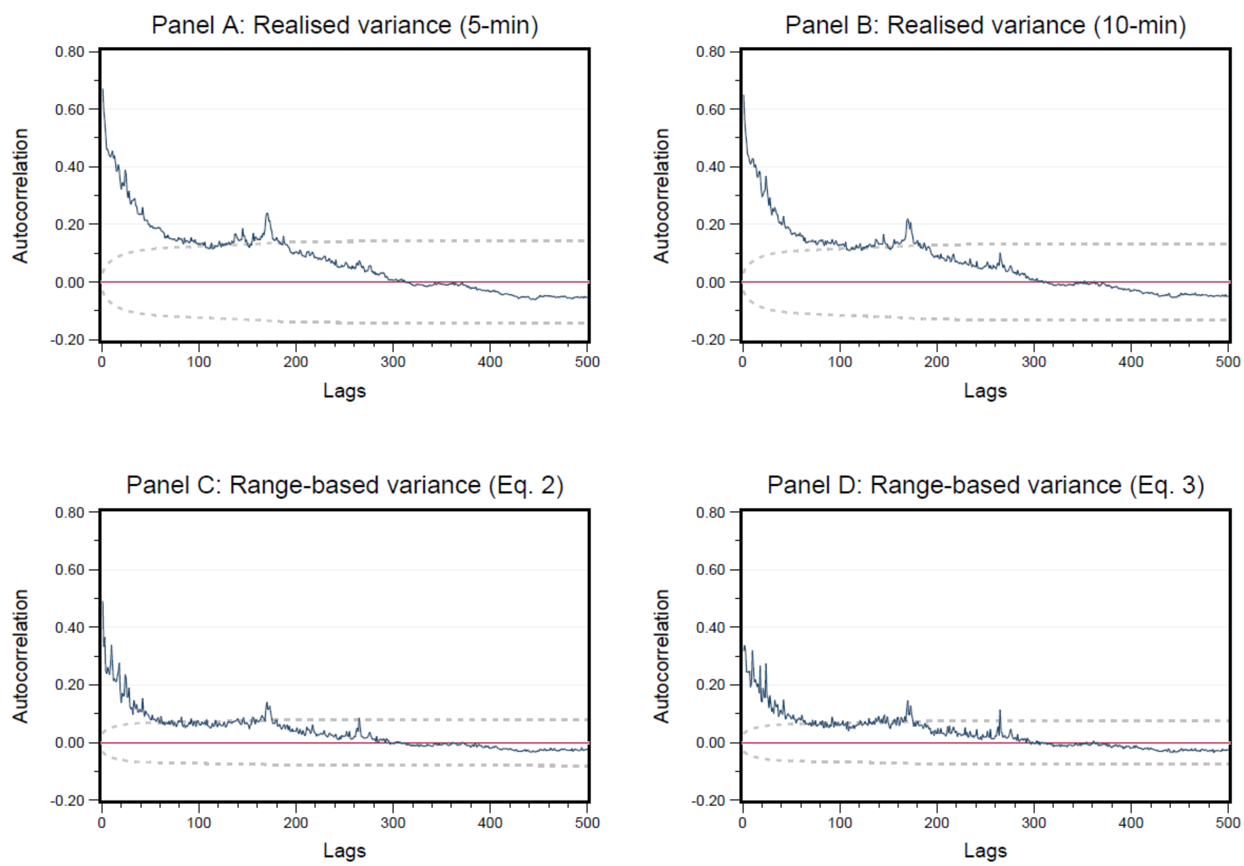

Figure C4. – In this figure, correlograms of the four different volatility proxies of the HSI index are demonstrated in the above panels. A window of 500 lags is used in all four panels. Furthermore, the light grey dashes represent the Bartlett's formula for $MA(q)$ 95% confidence bands, i.e. 5% significance limits.

Figure C5. – Autocorrelations of IPC volatility proxies

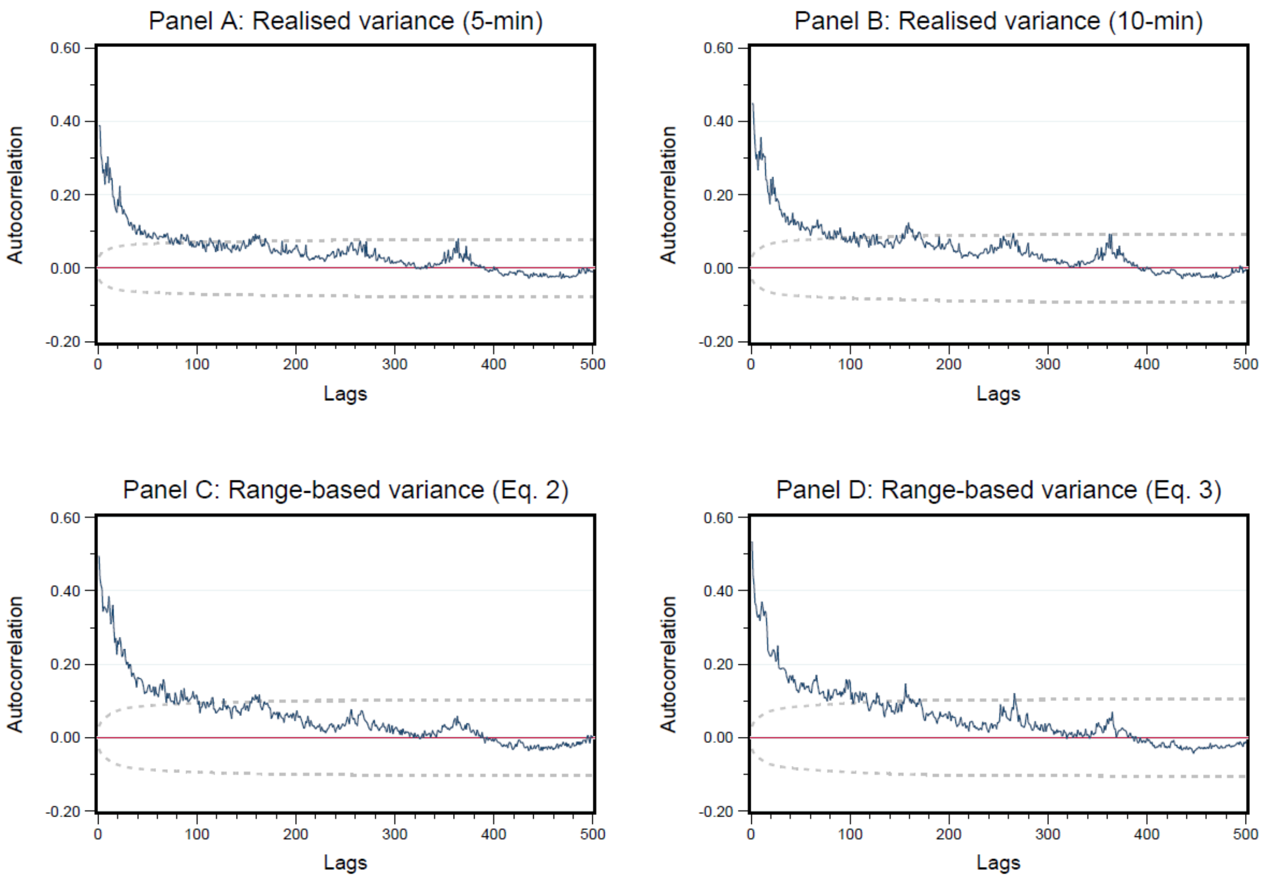

Figure C5. – In this figure, correlograms of the four different volatility proxies of the IPC index are demonstrated in the above panels. A window of 500 lags is used in all four panels. Furthermore, the light grey dashes represent the Bartlett's formula for $MA(q)$ 95% confidence bands, i.e. 5% significance limits.

Figure C6. – Autocorrelations of KOSPI volatility proxies

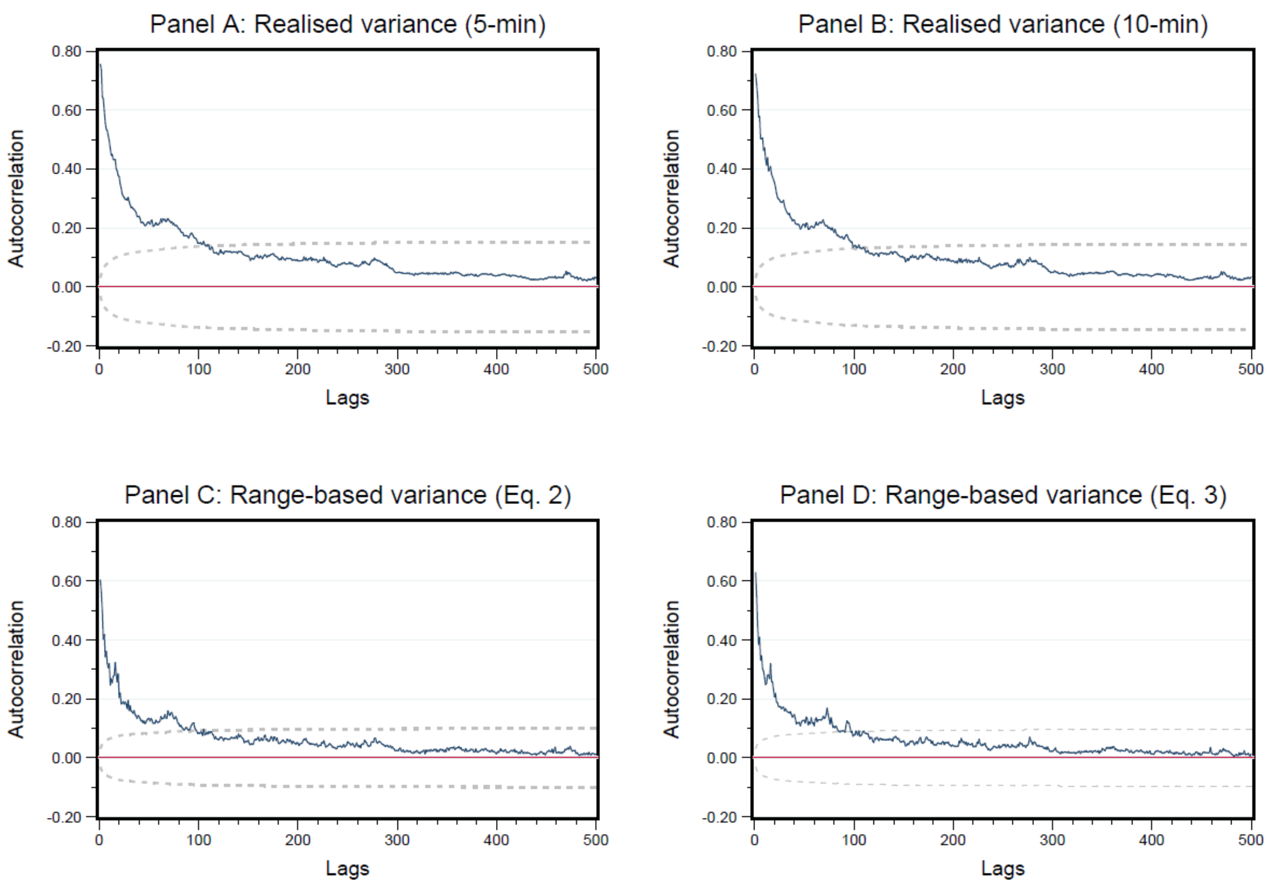

Figure C6. – In this figure, correlograms of the four different volatility proxies of the KOSPI index are demonstrated in the above panels. A window of 500 lags is used in all four panels. Furthermore, the light grey dashes represent the Bartlett's formula for $MA(q)$ 95% confidence bands, i.e. 5% significance limits.

Figure C7. – Partial autocorrelations of HSI volatility proxies

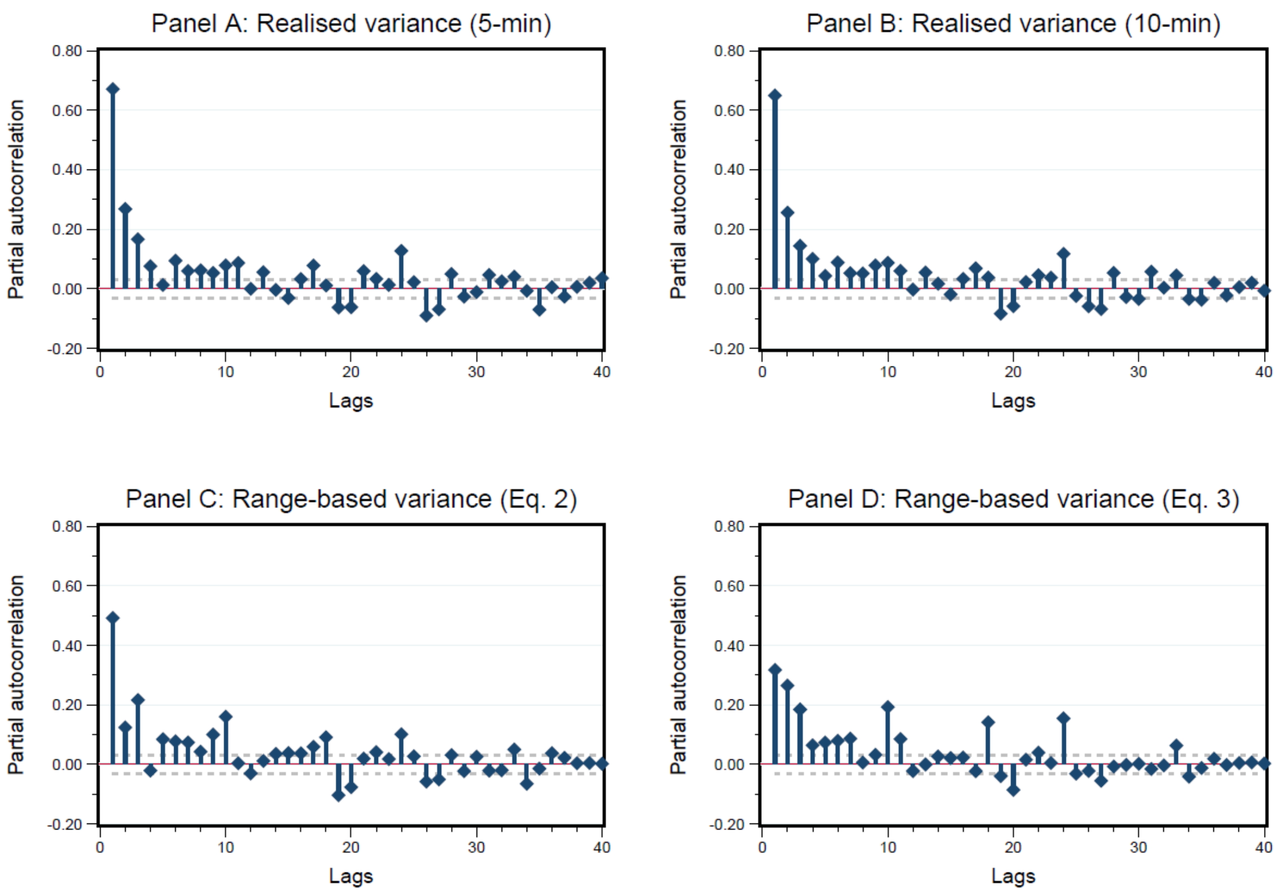

Figure C7. – In this figure, partial autocorrelations of the four different volatility proxies of the HSI index are demonstrated in the above panels. A window of 40 lags is used in all four panels. Furthermore, the light grey dashes represent the 95% confidence bands, which can be calculated as $se = 1/\sqrt{n}$.

Figure C8. – Partial autocorrelations of IPC volatility proxies

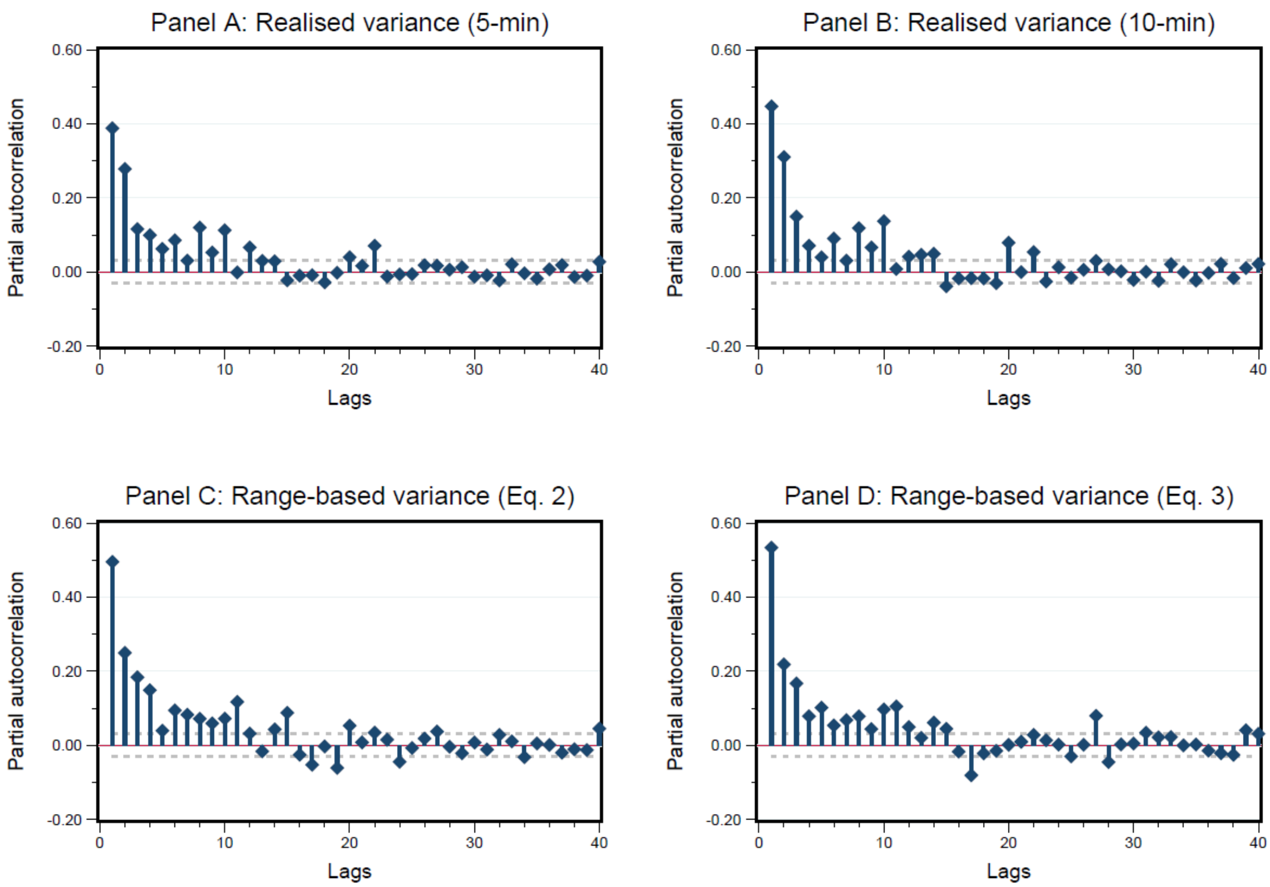

Figure C8. – In this figure, partial autocorrelations of the four different volatility proxies of the IPC index are demonstrated in the above panels. A window of 40 lags is used in all four panels. Furthermore, the light grey dashes represent the 95% confidence bands, which can be calculated as $se = 1/\sqrt{n}$.

Figure C9. – Partial autocorrelations of KOSPI volatility proxies

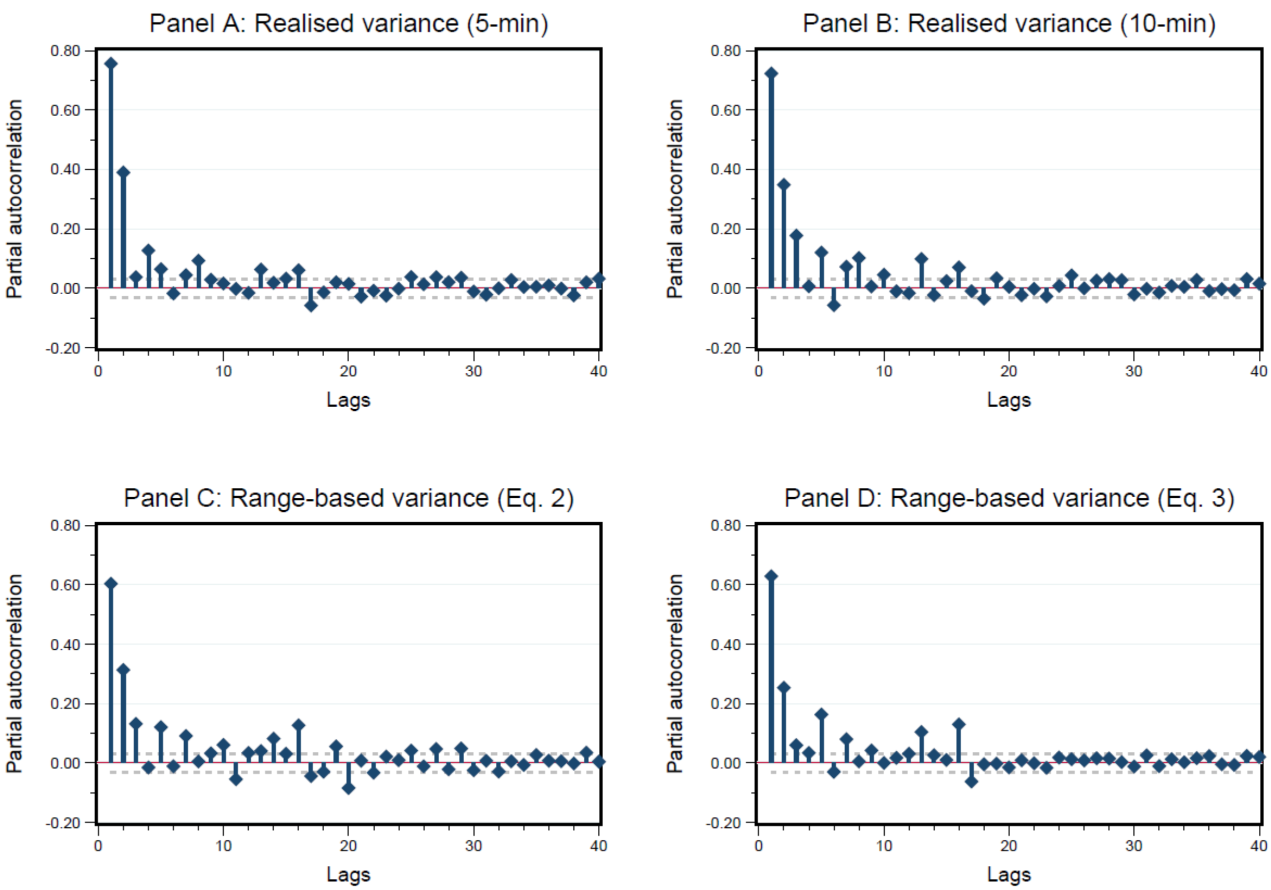

Figure C9. – In this figure, partial autocorrelations of the four different volatility proxies of the KOSPI index are demonstrated in the above panels. A window of 40 lags is used in all four panels. Furthermore, the light grey dashes represent the 95% confidence bands, which can be calculated as $se = 1/\sqrt{n}$.

Figure C10. – HSI index

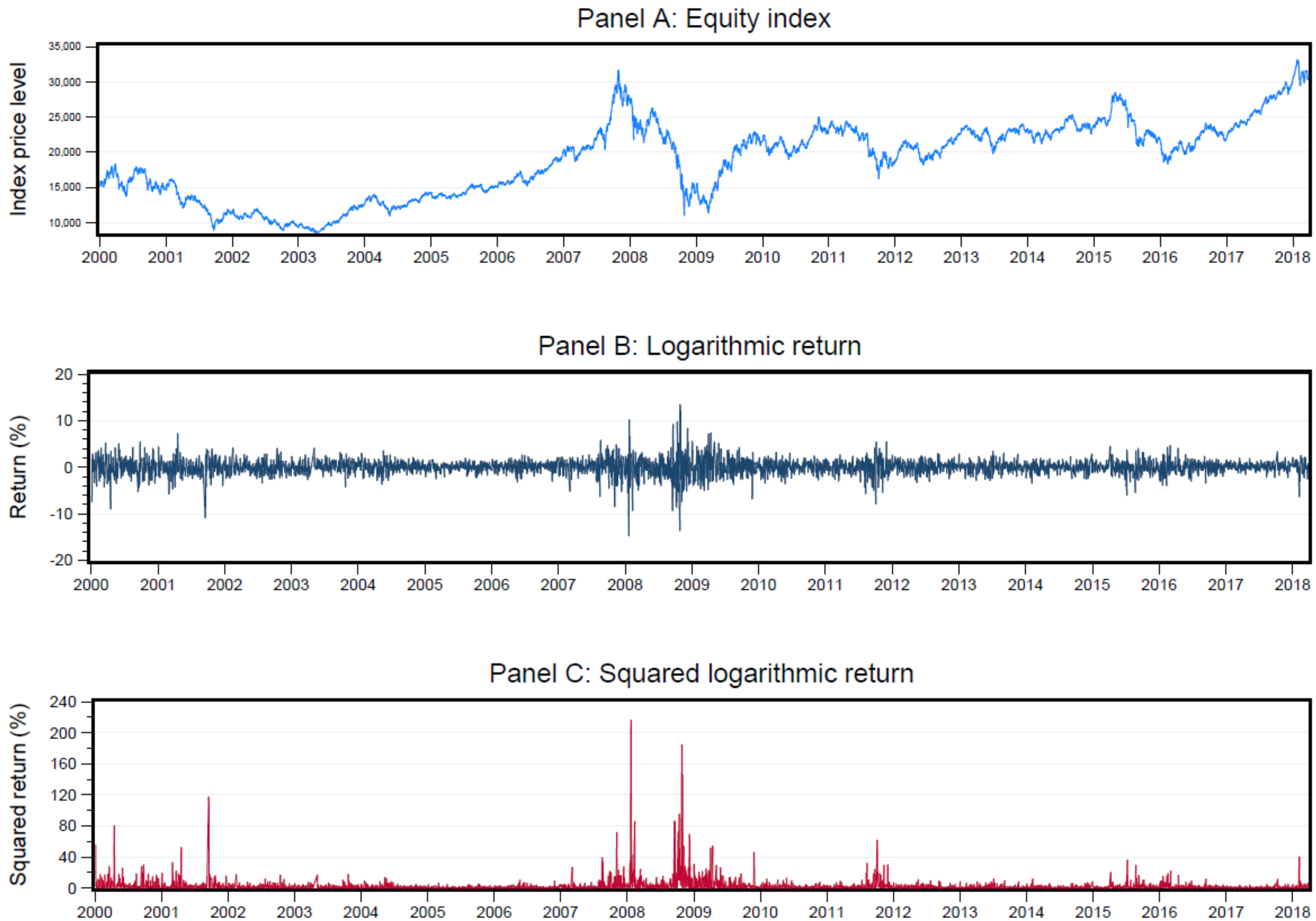

Figure C10. – This figure illustrates the historical data of the HSI index price level, log returns, and squared returns. These are displayed in **Panels A**, **B**, and **C**, respectively. The sample period starts from January 5, 2000, to March 27, 2018, which contains 4,078 observations.

Figure C11. – IPC index

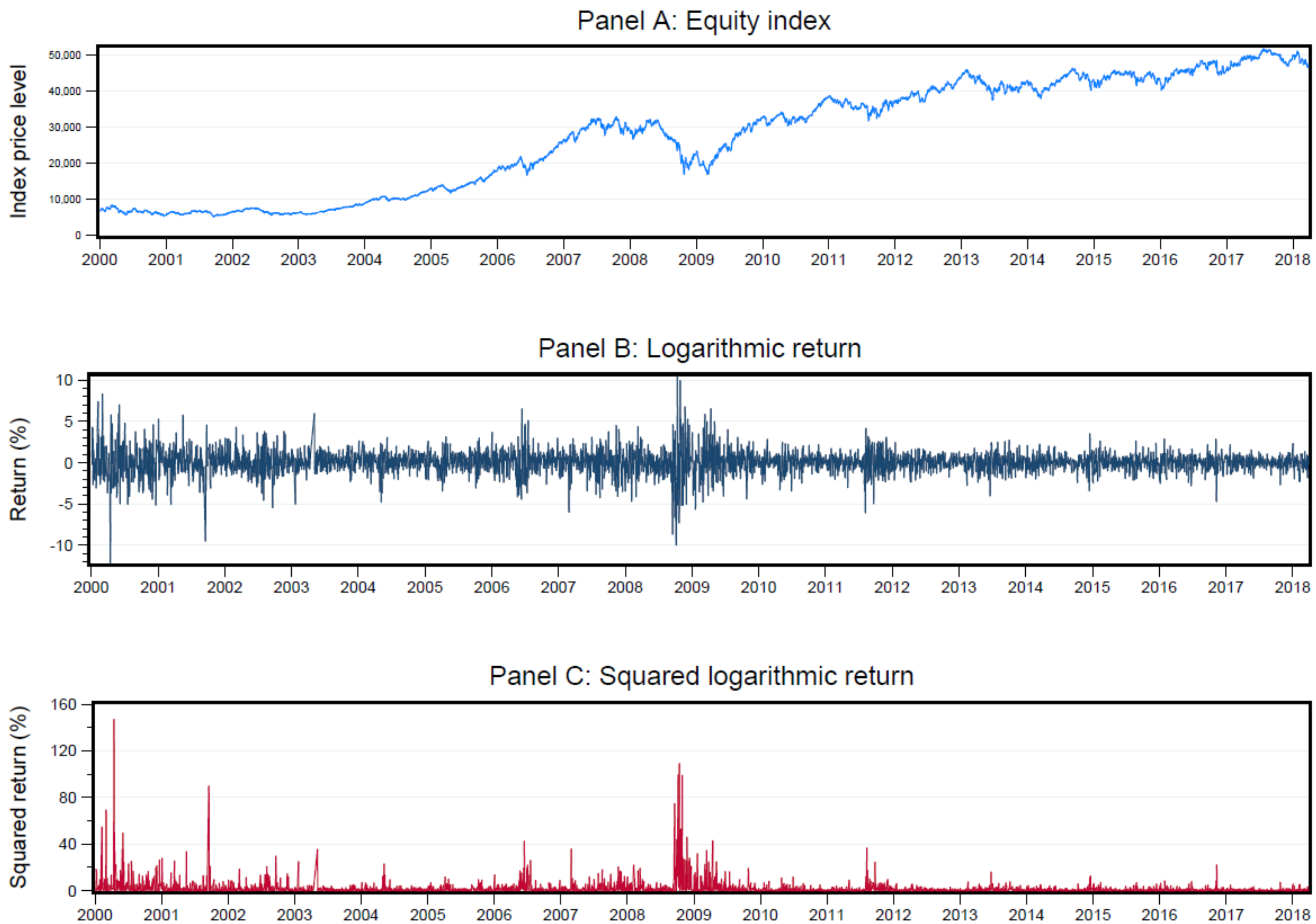

Figure C11. – This figure illustrates the historical data of the IPC index price level, log returns, and squared returns. These are displayed in **Panels A**, **B**, and **C**, respectively. The sample period starts from January 5, 2000, to March 27, 2018, which contains 4,078 observations.

Figure C12. – KOSPI index

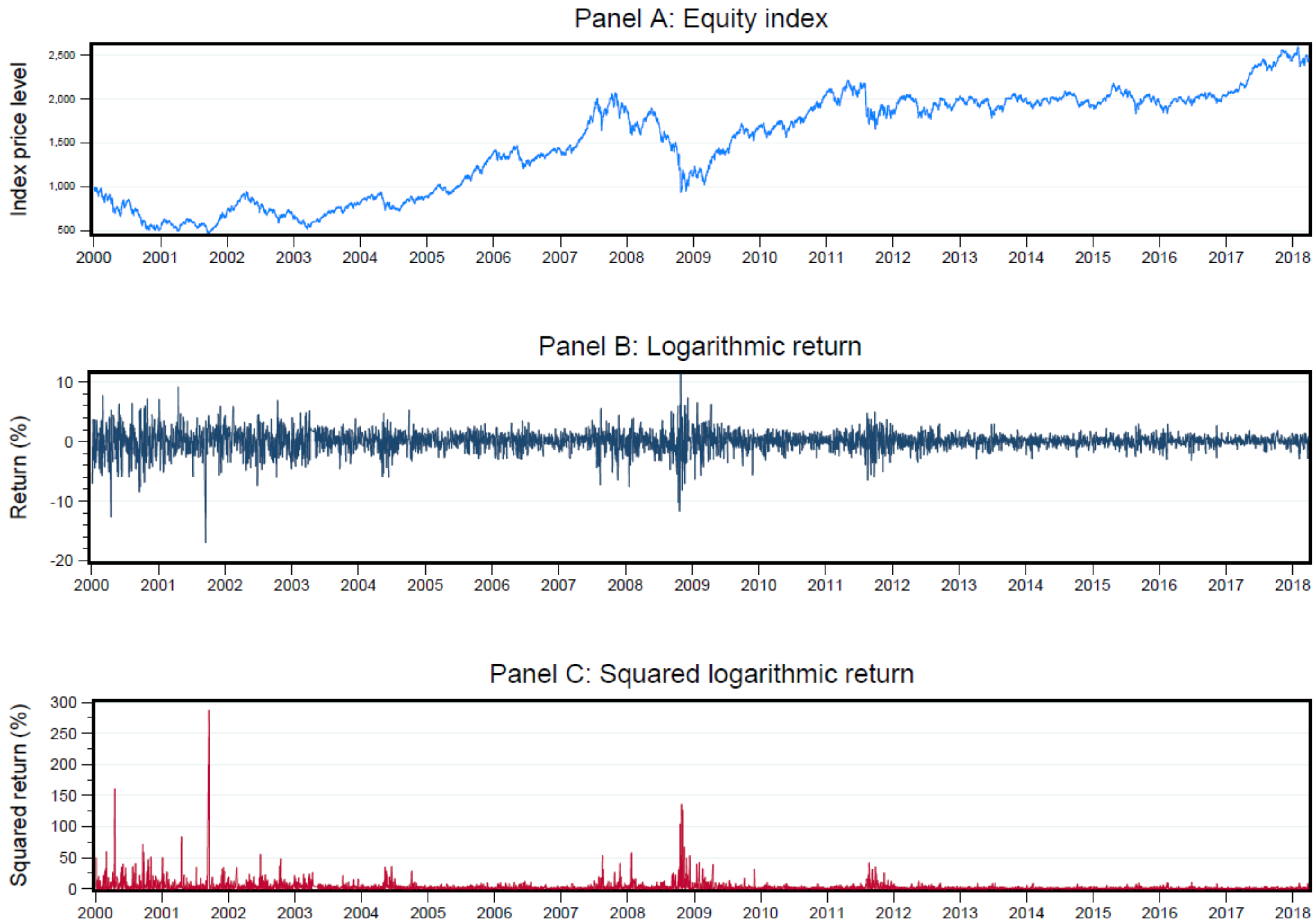

Figure C12. – This figure illustrates the historical data of the KOSPI index price level, log returns, and squared returns. These are displayed in **Panels A**, **B**, and **C**, respectively. The sample period starts from January 5, 2000, to March 27, 2018, which contains 4,078 observations.

Figure C13. – Historical behaviour of variance proxies of S&P 500

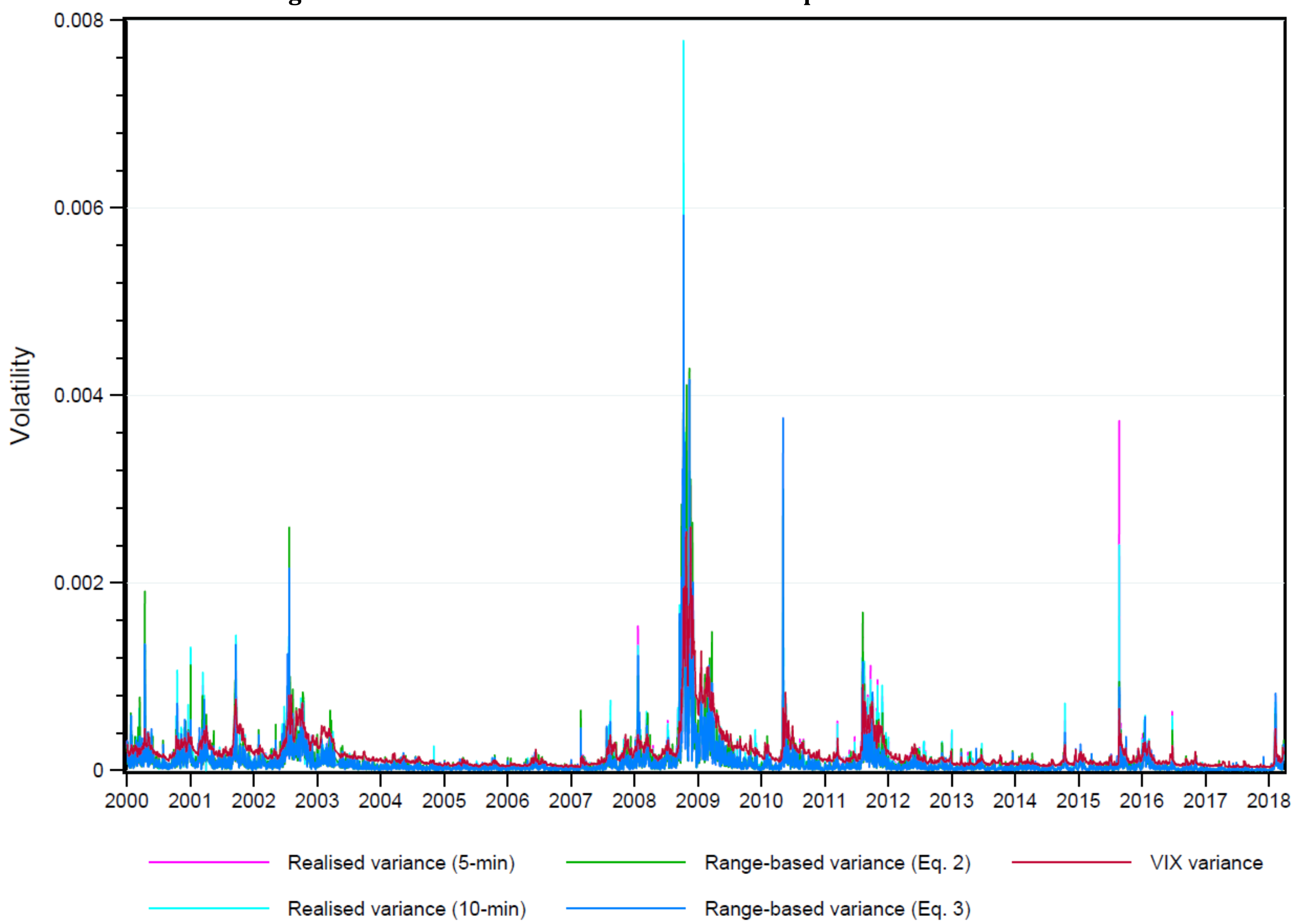

Figure C13. – This graph plot illustrates the historical behaviour of the variance proxies of the S&P 500. We observe that the VIX variance proxy, which is calculated as $\zeta_t = (1/252)(VIX_t/100)^2$, tracks other variance proxies rather well. The sample period starts from January 5, 2000, to March 27, 2018, which contains 4,078 observations.