

Recurrent Conditional Heteroskedasticity*

T.-N. Nguyen[†]

M.-N. Tran^{*}

R. Kohn[‡]

January 25, 2022

Abstract

We propose a new class of financial volatility models, called the REcurrent Conditional Heteroskedastic (RECH) models, to improve both in-sample analysis and out-of-sample forecasting of the traditional conditional heteroskedastic models. In particular, we incorporate auxiliary deterministic processes, governed by recurrent neural networks, into the conditional variance of the traditional conditional heteroskedastic models, e.g. GARCH-type models, to flexibly capture the dynamics of the underlying volatility. RECH models can detect interesting effects in financial volatility overlooked by the existing conditional heteroskedastic models such as the GARCH, GJR and EGARCH. The new models often have good out-of-sample forecasts while still explaining well the stylized facts of financial volatility by retaining the well-established features of econometric GARCH-type models. These properties are illustrated through simulation studies and applications to thirty-one stock indices and exchange rate data. An user-friendly software package, together with the examples reported in the paper, is available at <https://github.com/vbayeslab>.

Keywords. Deep Learning, volatility modelling, neural networks, conditional heteroskedasticity.

1 Introduction

Financial time series, e.g. currency exchange rates or stock returns, exhibit stylized facts such as volatility clustering and leverage effects. The volatility clustering phenomenon of financial time series refers to the observation that “large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes” (Mandelbrot, 1967). This behavior implies that the volatilities, i.e. the conditional standard deviations, of financial returns are observed to be highly autocorrelated and exhibit periods of both low and high volatility. The leverage effects exhibited in financial time series, on the other hand, relate to the observation that the negative and positive past returns have asymmetric effects

* We would like to thank the editor and the three anonymous referees for their constructive comments and suggestions.

[†] Discipline of Business Analytics, The University of Sydney Business School.

[‡] School of Economics, UNSW Business School. The research of Nguyen and Kohn was partially supported by the Australian Research Council’s Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS). Tran is partially supported by the ARC Discovery Project DP200103015.

on the volatility (Black, 1976). More specifically, the current volatility tends to be larger following a previous negative shock, i.e. a return below its expected value, than a positive one of the same absolute value. The volatility clustering and leverage imply that the volatility of financial assets changes over time, i.e., being heteroskedastic.

Time-varying volatility is a key assumption in the volatility modeling literature. A large number of volatility models have been developed since Engle (1982) proposed the Autoregressive Conditional Heteroskedastic (ARCH) model, which allows the conditional variance, i.e. the squared volatility, to change over time as a *deterministic* function of the historical shocks while leaving the unconditional variance unchanged. The most successful extension of the ARCH model is the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model of Bollerslev (1986); it models the current conditional variance as a linear function of the past conditional variances and squared returns. The GARCH model together with the ARCH model and their variants define a class of models, often referred to as the conditional heteroskedastic or GARCH-type models, which use deterministic functions of historical information, e.g. the past returns and past conditional variances, to model the current conditional variance and are able to capture the stylized facts of financial time series. Another notable line of research in the volatility modeling literature focuses on the stochastic volatility (SV) model (Taylor, 1986) and its variants, which formulate the conditional variance using latent *stochastic* processes that do not directly involve the past returns. Our article is, however, interested in GARCH-type models which are probably more popular in the volatility modeling literature because it is much easier to estimate GARCH-type models than SV-type alternatives. See Koopman et al. (2016) for a comprehensive comparison between the GARCH-type and SV-type models.

Recurrent neural networks (RNNs) in the Deep Learning literature are successfully used in a large number of industrial-level applications; e.g. language translation, image captioning, speech synthesis. RNNs are well-known for their ability to efficiently capture the long-range memory and non-linear serial dependence existing within various types of sequential data, and are considered as the state-of-the-art models for many sequence learning problems (Lipton et al., 2015). See Goodfellow et al. (2016) for a comprehensive discussion of various types of neural network models (NNs) and their broad range of applications. The recent success of RNNs has motivated econometricians to incorporate RNNs and other deep learning models into their econometric models. There is a large amount of research along this line, with a focus on modelling the mean rather than the variance of financial asset returns; see, e.g., Zhang et al. (1998); Zhang (2003). Leveraging the power of deep learning in volatility modelling is still somewhat overlooked in the econometrics literature. Donaldson and Kamstra (1997) are one of the first to propose a NN-GARCH model that adds a feedforward NN component into the conditional variance of the GJR model (Glosten et al., 1993). Their in-sample and out-of-sample results on four stock markets suggest that the NN-GARCH model is preferred to several benchmark GARCH-type models. Roh (2007) proposes NN-based volatility models that first estimate the conditional variance by an econometric volatility model, then use these estimates as inputs to a feedforward neural network, which then non-linearly transforms these inputs to output the final estimate of the conditional variance. Kim and Won (2018) extend this idea by using the outputs from several GARCH-type models rather than a single one as the inputs to a recurrent neural network; see also Luo et al. (2018). Liu (2019) uses the Long Short-term Memory, a sophisticated RNN technique, for volatility modelling and reports

its improved prediction over GARCH in two datasets. In general, these hybrid models that combine neural networks and financial econometric models are empirically superior to several econometric models and plain neural network models, in terms of predictive performance. However, these models are often engineering-oriented and ignore the interpretable aspects of econometrics volatility models and the important stylized facts of financial time series. Some of these existing models use feedforward NNs rather than RNNs and thus might ignore the time effects in time series data. Their design is also rather inflexible in the sense that they combine a particular econometric model with a particular NN. It is important to design a more flexible framework that is easy to adapt to advances in both the deep learning and volatility modelling literature.

GARCH-type models are generally simple yet highly interpretable in the sense that they are designed to explain the distinct behaviors of financial time series. Any new volatility model should not overlook this interpretability of the traditional econometric models. This paper proposes a new class of models, called the REcurrent Conditional Heteroskedastic models (RECH), that not only improve the forecast performance of GARCH-type models, by leveraging the capability of learning non-linearity and long-range dependence of RNNs, but also place significant emphasis on the interpretation of the estimated volatility, by inheriting the well-established features of GARCH-type models. We now briefly explain why RECH models fit well within the volatility modeling literature. First, similarly to GARCH-type models, the conditional variances in RECH models are a deterministic function of the past values; hence it is easy to estimate RECH models as their likelihood functions can be evaluated analytically. Second, RECH models are still able to explain the stylized facts of the underlying volatility dynamics. Third, by inheriting the predictive power from deep learning techniques, RECH models often forecast better. Fourth, the highly flexible design of RECH models makes it easy to adopt advances in both the deep learning and volatility modeling literatures, allowing it to be used in a wide range of applications in financial time series analysis. A Matlab software package implementing Bayesian estimation and inference for RECH models together with the examples reported in this paper is available at <https://github.com/vbayeslab>.

The rest of the article is organized as follows. Section 2 briefly reviews the GARCH model and its variants. Section 3 briefly reviews different types of neural networks and proposes RECH models. Section 4 discusses Bayesian estimation and inference for RECH models. Section 5 presents the simulation study and applies RECH models to analyze four benchmark financial datasets. Section 6 concludes. The Appendix gives implementation details and further empirical results.

2 Conditional heteroscedastic models

Let $y = \{y_t, t = 1, \dots, T\}$ be a time series of demeaned returns and \mathcal{F}_t be the σ -field of the information up to time t . Conditional heteroskedastic models represent the conditional variance $\sigma_t^2 := \text{Var}(y_t | \mathcal{F}_{t-1})$ of the observation y_t as a deterministic function of the observations and the conditional variances in the previous time steps. Mathematically, these models are expressed as:

$$y_t = \sigma_t \epsilon_t, \quad \epsilon_t \sim i.i.d \quad \text{with } t = 1, 2, \dots, T, \quad (1a)$$

$$\sigma_t^2 = \omega + f(\sigma_{t-1}^2, \dots, \sigma_{t-p}^2, y_{t-1}, \dots, y_{t-q}, \theta); \quad (1b)$$

$f(\cdot)$ is a positive deterministic function parameterized by the vector of unknown parameters θ ; $p, q \geq 0$ are the number of lags of σ_t^2 and y_t respectively; and ω is a non-negative constant ensuring that the conditional variance σ_t^2 is positive. The shocks ϵ_t are i.i.d. with zero mean and unit variance.

The GARCH model (Bollerslev, 1986) formulates the conditional variance σ_t^2 as a linear combination of the previous returns and conditional variances in a ARMA(p, q) form as:

$$y_t = \sigma_t \epsilon_t, \quad t = 1, 2, \dots, T, \quad (2a)$$

$$\sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i y_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2, \quad t = 2, \dots, T; \quad (2b)$$

$\omega > 0, \alpha_i, \beta_j \geq 0, i = 1, \dots, p, j = 1, \dots, q$ and $\sum_{i=1}^p \alpha_i + \sum_{j=1}^q \beta_j < 1$ to ensure the stationarity of the GARCH process.

The structure of the conditional variance in (2b) is symmetric in the sense that the conditional variance σ_t^2 does not depend on the sign of the y_t , implying that the conventional GARCH(p, q) model cannot capture the important leverage effect, i.e., σ_t^2 depends asymmetrically on the previous returns, in financial time series. Glosten et al. (1993) propose a variant of the GARCH model, often called the GJR model, of the form

$$y_t = \sigma_t \epsilon_t, \quad t = 1, 2, \dots, T, \quad (3a)$$

$$\sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i y_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 + \sum_{i=1}^p \gamma_i \mathbb{1}[y_{t-i} < 0] y_{t-i}^2; \quad (3b)$$

$\omega > 0, \alpha_i, \beta_j \geq 0, \alpha_i + \gamma_i \geq 0, i = 1, \dots, p, j = 1, \dots, q$ and $\sum_{i=1}^p \alpha_i + \sum_{j=1}^q \beta_j + \sum_{i=1}^p \gamma_i < 1$ to ensure the stationarity of the y_t process and the positivity of the conditional variance. The indicator function $\mathbb{1}[y_t < 0]$ in (3b) equals 1 if $y_t < 0$ and is 0 otherwise. In the GJR model, if $\gamma_i > 0$, negative returns are more influential than positive returns.

Another popular GARCH-type model is the Exponential Garch (EGARCH) model of Nelson (1991)

$$y_t = \sigma_t \epsilon_t, \quad t = 1, 2, \dots, T, \quad (4a)$$

$$\log \sigma_t^2 = \omega + \sum_{i=1}^p \beta_i \log \sigma_{t-i}^2 + \sum_{j=1}^q \alpha_j \left[\frac{|y_{t-j}|}{\sigma_{t-j}} - \text{E} \left\{ \frac{|y_{t-j}|}{\sigma_{t-j}} \right\} \right] + \sum_{j=1}^q \gamma_j \frac{y_{t-j}}{\sigma_{t-j}}, \quad (4b)$$

where the roots of the polynomial $(1 - \beta_1 L - \dots - \beta_p L^p)$ must lie outside the unit circle to ensure the stationarity of the y_t process. By working on the log-scale, the EGARCH model removes the positivity constraints on the model parameters and states the leverage terms in Eq. (4b) to capture the asymmetry in volatility clustering. See Poon and Granger (2003) and Bollerslev (2008) for a comprehensive discussion of the family of GARCH models and their properties. We will use GARCH, GJR and EGARCH as the benchmark econometric models to compare against RECH models, because they are widely used in the volatility modelling literature.

3 Recurrent conditional heteroskedastic models

3.1 Recurrent neural network models

This section denotes the time series data as $\{D_t = (x_t, z_t), t = 1, 2, \dots\}$, where $x_t = (x_{t,1}, \dots, x_{t,K})^\top$ is the vector of inputs and z_t the scalar output. For the sequence $\{x_t\}$, $x_{i:j}$ denotes (x_i, \dots, x_j) for $i \leq j$. The goal of recurrent neural network models is to model the conditional distribution $p(z_t|x_t, D_{1:t-1})$.

There are several standard time series models. One approach is to represent time effects *explicitly* via some simple functions, often a linear function, of the lagged values of the time series. This is the mainstream time series data analysis approach with the well-known ARIMA method (Box and Jenkins, 1976). This section considers an alternative approach representing time effects *implicitly* via latent variables that are designed to store the memory of the dynamics in the data. These latent variables, also called hidden states, are updated recurrently using the information carried over by their values from the past and the information from the data at the current time. Recurrent neural networks (RNNs), belonging to the second category, were first developed in cognitive science (Elman, 1990) and successfully used in machine learning.

If the serial dependence structure is ignored, then a feedforward neural network (FNN) can be used to transform the raw input data x_t into a set of hidden units h_t , often called *learned features*, for the purpose of explaining or predicting z_t . Figure 1 is a graphical representation of a FNN model with one hidden layer containing L hidden units.

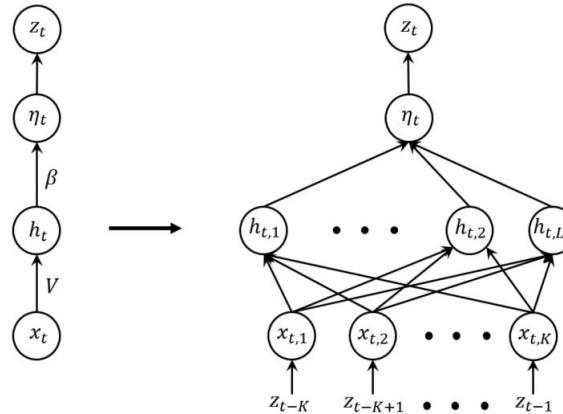


Figure 1: A FNN model with one hidden layer in compact (*left*) and explicit (*right*) styles.

Given the FNN model in Figure 1, the output z_t is calculated as:

$$h_t = \phi(Vx_t + b), \quad (5a)$$

$$\eta_t = \beta^\top h_t + \beta_0, \quad (5b)$$

$$z_t|\eta_t \sim p(z_t|\eta_t); \quad (5c)$$

V is a $L \times K$ matrix of weights connecting the input layer to the hidden layer; $\beta = (\beta_1, \dots, \beta_L)^\top$ is a vector of weights connecting the hidden layer to the output layer; β_0 is a scalar; $b =$

$(b_1, \dots, b_L)^\top$ is a bias vector and $\phi(\cdot)$ is a non-linear scalar function, called the activation function. The scalar function ϕ is applied component-wise to a vector. In modern neural network modelling, the default recommendation for $\phi(\cdot)$ is to use the rectified linear unit (Nair and Hinton, 2010; Le et al., 2015), or ReLU, having the form $\phi(z) = \max\{0, z\}$. The density $p(z_t|\eta_t)$ depends on the learning task. For example, if z_t is continuous, then typically $p(z_t|\eta_t)$ is a normal distribution with mean η_t and variance σ^2 ; if z_t is binary, then $z_t|\eta_t$ follows a Bernoulli distribution with probability $\text{logit}^{-1}(\eta_t)$.

FNNs provide a powerful way to approximate the true function that maps the input x_t to the mean $E(z_t|x_t)$ or to transform the raw data x_t into summary statistics h_t having some desirable properties. However, FNNs are unsuitable for time series data analysis as the time effects and the serial correlations are ignored. The main idea behind RNNs is to let the set of hidden units h_t feed itself on its lagged value h_{t-1} . Hence, an RNN can be best thought of as a FNN that allows a connection of the hidden units to their value from the previous time step, enabling the network to possess memory. This basic RNN model (Elman, 1990) can be written as:

$$h_t = \phi(Vx_t + Wh_{t-1} + b), \quad (6a)$$

$$\eta_t = \beta^\top h_t + \beta_0, \quad (6b)$$

$$z_t|\eta_t \sim p(z_t|\eta_t); \quad (6c)$$

the parameters are the bias vector b , the bias scalar β_0 , the weight matrices V , W , and β for input-to-hidden, hidden-to-hidden and hidden-to-output connections, respectively. Similarly to FNNs, $\phi(\cdot)$ is a non-linear activation function; common choices are the ReLU or the sigmoid $\phi(z) = 1/(1 + e^{-z})$. Usually, we can set $h_1 = 0$, i.e. the neural network initially memoryless.

Figure 2 (Nguyen et al., 2019) graphically illustrates the RNN model (6a)-(6c). The circuit diagram (*Left*) can be interpreted as an unfolded computational graph (*Right*), where each node is associated with a particular time step. The calculation of h_t can be represented as a Simple Recurrent Neuron (SRN) unit, as Figure 3 shows, and we refer to (6a) as $h_t = \text{SRN}(x_t, h_{t-1})$, taking data x_t at time t and the previous state h_{t-1} as the inputs. Using the SRN structure, the unfolded graph of the RNN model of Elman (1990), which is normally referred to as the Simple RNN model, can be reinterpreted as the unfolded graph in Figure 3(*right*).

There are more sophisticated recurrent neuron unit structures to compute h_t , such as the memory cell in the Long Short-Term Memory (LSTM) model of Hochreiter and Schmidhuber (1997) and the Statistical Recurrent Unit (SRU) of Oliva et al. (2017). Below, we sometimes write $h_t = \text{RNN}(x_t, h_{t-1})$ to represent the way the hidden state h_t is computed in an RNN model. Our article, however, only considers the SRN.

3.2 Recurrent conditional heteroskedastic models

3.2.1 The general formulation

The key motivation of RECH models is to allow the constant term ω in the general formulation of GARCH-type models in (1a)-(1b) to be driven by an auxiliary deterministic process governed by an RNN, in order to capture complex dynamics such as non-linearity and long-term dependence that might not be captured efficiently by the GARCH-type component

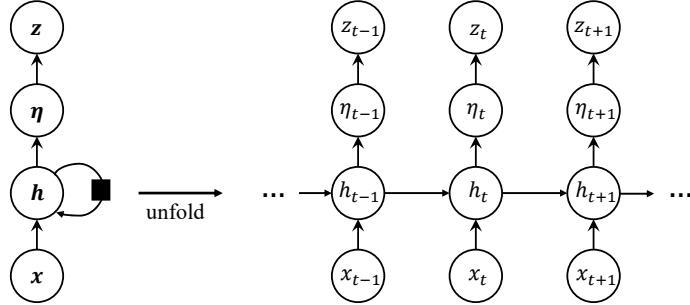


Figure 2: Graphical representation of the RNN model in (6a)-(6c).

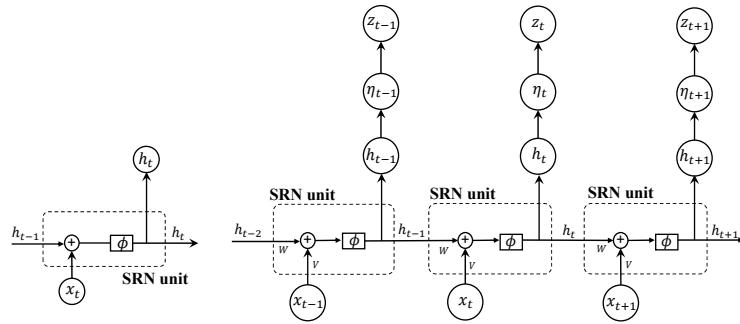


Figure 3: The structures of the SRN unit (*left*) and the graphical representation of the Simple RNN model (*right*), which uses the SRN unit to compute the latent state h_t . The \oplus symbols represents the addition operation.

$f(\sigma_{t-1}^2, \dots, y_{t-q})$. Our general RECH model is written as:

$$y_t = \sigma_t \epsilon_t, \quad \epsilon_t \sim i.i.d., \quad t = 1, 2, \dots, T \quad (7a)$$

$$\sigma_t^2 = g(\omega_t) + f(\sigma_{t-1}^2, \dots, \sigma_{t-p}^2, y_{t-1}, \dots, y_{t-q}), \quad t = 2, \dots, T, \quad \sigma_1^2 = \sigma_0^2 \quad (7b)$$

$$\omega_t = \beta^\top h_t + \beta_0, \quad t = 2, \dots, T, \quad (7c)$$

$$h_t = \text{RNN}(x_t, h_{t-1}), \quad t = 2, \dots, T, \quad \text{with } h_1 \equiv 0; \quad (7d)$$

$g(\cdot)$ is a non-negative activation function; p and q are lag orders of σ_t^2 and y_t respectively; β_0 is a scalar; $\beta = (\beta_1, \dots, \beta_L)$ is the weight vector with L the number of hidden states. The reason an activation function is applied to ω_t is to ensure the conditional variance σ_t^2 is positive. We refer to ω_t in (7b) as the *recurrent* component, as it is driven by a RNN, and $f(\cdot)$ as the *GARCH* component as this is formed based on the GARCH-type structures without the constant term. Hence, we shall refer to the parameters of the recurrent component as the recurrent parameters, and refer to the parameters in $f(\cdot)$ as the GARCH parameters. The recurrent state $h_t = \text{RNN}(x_t, h_{t-1})$ takes as its inputs the previous state h_{t-1} and a vector of additional information x_t whose choice is discussed shortly.

The conditional variance in (7b) is a sum of the recurrent and the GARCH components.

This flexible design allows RECH models to enjoy many advances from both worlds of deep learning and volatility modelling. Similarly to deep learning models, RECH models can use highly sophisticated neural network structures to capture complicated dynamics, e.g., long-range dependence and non-linearity, of the volatility dynamics and hence improve the forecasting of traditional GARCH-type models in applications where the underlying volatility dynamics exhibits long memory and nonlinearity. Similarly to GARCH-type models, RECH models use simple yet interpretable structures to simulate important stylized facts in financial time series such as volatility clustering and leverage effects. RECH models are well suited to modeling volatility because they inherit many properties from the GARCH-type models and distinguish themselves from the existing NN-based volatility models that often overlook the interpretability of the mainstream econometric models. As the recurrent and GARCH components are additive, increasing the complexity of the recurrent component ω_t will not decrease the interpretability of the GARCH component, and hence of RECH models. The general formulation in (7a)-(7d) implies that most (if not all) variants of the GARCH models are nested in the RECH framework, because RECH models reduce to the corresponding GARCH-type models if $\beta = 0$.

In previous work that combined an NN-based component with a GARCH-type model, Donaldson and Kamstra (1997) added a FNN-based component to a GJR model and reported some improvement of their FNN-GJR model compared to several benchmark GARCH-type models. However, as discussed above, it might be inefficient to use a FNN to analyze time series data as FNNs are typically designed for cross-sectional data. Also, the estimation method of Donaldson and Kamstra (1997) uses randomized weights for the FNN component rather than optimizing them; this technique is not recommended in the modern deep learning literature (Goodfellow et al., 2016). Nguyen et al. (2019) incorporate LSTM into the stochastic volatility models and name their model LSTM-SV. LSTM-SV belongs to the class of parameter-driven models while RECH is an observation-driven model; see Koopman et al. (2016) for a comprehensive comparison between these two classes of models. More specifically, the volatility dynamics in RECH is a *deterministic* process rather than a stochastic latent process as in LSTM-SV, making it much easier to estimate RECH models than the LSTM-SV model as the likelihood of RECH models can be calculated analytically. The RNN component of the LSTM-SV model takes only its past values as inputs while the recurrent component ω_t of RECH models allows any information including past observations as inputs. We show below that this flexibility enables RECH models with SRN to capture complicated dynamics and the leverage effect in financial time series without needing to use complicated RNN structures such as LSTM. Finally, like GARCH and SV, RECH and LSTM-SV complement each other and offer different perspectives towards the volatility modelling problem. The Appendix compares the performance of the RECH and LSTM-SV models.

3.2.2 Specifications for RECH models

The RECH framework is highly flexible because it can easily incorporate advances from both the deep learning and volatility modeling literatures to design the recurrent and GARCH components, respectively. For example, by using the SRN structure for the recurrent component ω_t and the conditional variance structure of the GARCH(1,1) model for the GARCH

component, we obtain the SRN-GARCH specification of RECH model as:

$$y_t = \sigma_t \epsilon_t, \quad \epsilon_t \stackrel{iid}{\sim} \mathcal{N}(0, 1), \quad t = 1, 2, \dots, T \quad (8a)$$

$$\sigma_t^2 = \omega_t + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2, \quad t = 2, \dots, T, \quad \sigma_1^2 = \sigma_0^2 \quad (8b)$$

$$\omega_t = \beta_0 + \beta_1 h_t, \quad t = 2, \dots, T, \quad (8c)$$

$$h_t = \text{SRN}(x_t, h_{t-1}), \quad t = 2, \dots, T, \quad \text{with } h_1 \equiv 0; \quad (8d)$$

x_t is the input vector of the RNN at time t . Figure 4 graphically represents the SRN-GARCH model. For simplicity, we consider the standard normal distribution $\mathcal{N}(0, 1)$ for the errors ϵ_t . Here, we have used a linear activation function for $g(\cdot)$ in Eq. (7b), and set $\beta_0, \beta_1 \geq 0$ to ensure the positivity of the conditional variance. Alternatively, one can use a positive activation function for $g(\cdot)$, such as the ReLu or sigmoid, and relax the positivity constraints for β_0 and β_1 . We follow the GARCH literature and put the stationarity and positivity constraints on the GARCH parameters α and β , i.e., $\alpha, \beta \geq 0$ and $\alpha + \beta < 1$. These constraints do not imply that the SRN-GARCH model is stationary, but might improve numerical stability in estimation.

The recurrent function SRN is

$$\text{SRN}(x_t, h_{t-1}) := \phi(v^\top x_t + wh_{t-1} + b),$$

with $\phi(\cdot)$ a non-linear activation function. We use the bounded ReLU activation for $\phi(\cdot)$ as it is easier to train and often performs better than other alternatives in the deep learning literature (Liew et al., 2016). The bounded activation also guarantees a finite unconditional volatility; see Theorem 1. The recurrent weight w and offset term b are scalars. By default, we use only one recurrent state, i.e. h_t is a scalar; however, it is possible to extend the specification in (8a)-(8d) to the case where h_t is a vector of hidden states. There is no restriction on the choice of the input vector x_t ; typically, x_t should include variables such as covariates and past returns that are deemed useful for predicting volatility σ_t^2 . If there are no covariates, as in the applications below, our choice for x_t is the vector of the past return y_{t-1} and the past conditional variance σ_{t-1}^2 . We also found it useful to include in x_t the past recurrent component ω_{t-1} . Hence, $x_t = (w_{t-1}, y_{t-1}, \sigma_{t-1}^2)^\top$ and the input weights are $v = (v_0, v_1, v_2)^\top$.

While a non-zero β in (8b) quantifies the linear dependence of the current conditional variance σ_t^2 on its past value σ_{t-1}^2 , a non-zero weight v_2 quantifies the non-linear dependence of σ_t^2 on σ_{t-1}^2 . That is, the recurrent component allows non-linear dependence of σ_t^2 on σ_{t-1}^2 . It is also well perceived in the deep learning literature that RNNs are able to capture long-range dependence (Greaves-Tunnell and Harchaoui, 2019), therefore the recurrent component can allow long-range dependence that $\{\sigma_s^2, s < t\}$ have on σ_t^2 .

We note that the SRN-GARCH specification uses the conditional variance structure of the GARCH model, but is still able to capture the leverage effects of the volatility dynamics as the input vector x_t includes the asymmetric leverage term y_{t-1} itself, not y_{t-1}^2 . Section 5 shows that this choice of the input vector x_t makes the volatility estimated by RECH models less sensitive to the choice of the structure for the GARCH component. For selecting the lags p and q in the GARCH component, we find that SRN-GARCH(1,1) often works well in almost all cases. This is probably because the recurrent component ω_t is able to capture the long-range dependence (Greaves-Tunnell and Harchaoui, 2019), hence larger lags are unnecessary. This is also consistent with the observation in the financial econometrics literature that the

GARCH(1,1) model often works the best among other GARCH models (Hansen and Lunde, 2005). Below, by SRN-GARCH without mentioning the lags we mean SRN-GARCH(1,1). We note that the SRN-GARCH specification simplifies to the GARCH(1,1) model if $\beta_0 > 0$ and $\beta_1 = 0$.

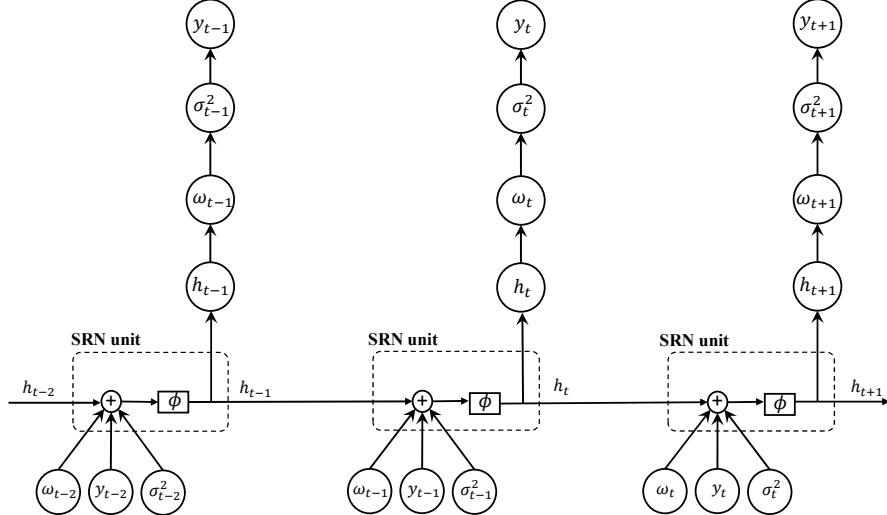


Figure 4: Graphical representation of SRN-GARCH.

Many other specifications of RECH models can be constructed. For example, the SRN-GJR specification is obtained by using the SRN structure for the recurrent component and the conditional variance structure of the GJR(1,1) for the GARCH component. Table 1 presents several RECH model specifications. It is also possible to use other RNN structures such as LSTM (Hochreiter and Schmidhuber, 1997) or SRU (Oliva et al., 2017) for the recurrent component. It is worth noting that the GJR and EGARCH models accommodate the leverage effects as linear terms in the conditional variance equation and hence can only capture the linear dependence of the leverage effects. RECH models, on the other hand, are able to capture other leverage dependence rather than the linearity, e.g. non-linearity or temporal dependence, of the leverage effects by allowing the leverage term y_{t-1} to be an input of the RNN.

Given the general formulation of RECH models in (7a)-(7d), its σ_t^2 process, and thus its y_t process, is not guaranteed to be stationary unless $\beta_1 = 0$ and the GARCH parameters satisfy the stationary constraints of the corresponding GARCH components $f(\cdot)$. For example, the SRN-GARCH specification in (8a)-(8d) is stationary if $\beta_0, \alpha, \beta > 0, \beta_1 = 0, \alpha + \beta < 1$. Although non-stationarity for volatility may be mathematically less appealing, it is often argued to be more realistic in practice, e.g. van Bellegem (2012). Theorem 1 below guarantees that the variance of y_t is bounded.

Theorem 1. *Consider the SRN-GARCH specification in (8a)-(8d). Assume that*

- $\alpha, \beta > 0, \alpha + \beta < 1$;
- *the recurrent component is bounded, i.e. $\omega_t \leq M$, almost surely for some $M < \infty$.*

Models	Conditional variance	Constraints
SRN-GJR	$\sigma_t^2 = \omega_t + \alpha y_{t-1}^2 + \beta \sigma_{t-1}^2 + \gamma I_{[y_{t-1}<0]} y_{t-1}^2$ $\omega_t = \beta_0 + \beta_1 h_t$ $h_t = \text{SRN}(x_t, h_{t-1}), \text{ with } x_t = (\omega_{t-1}, y_{t-1}, \sigma_{t-1}^2)$	$\alpha, \beta \geq 0$ $\alpha + \gamma \geq 0$ $\alpha + \beta + \gamma < 1$ $\beta_0, \beta_1 \geq 0$
SRN-EGARCH	$\sigma_t^2 = \omega_t + \exp \left\{ \omega + \beta \log \sigma_{t-1}^2 + \alpha \left[\frac{ y_{t-1} }{\sigma_{t-1}} - \sqrt{\frac{2}{\pi}} \right] + \gamma \frac{y_{t-1}}{\sigma_{t-1}} \right\}$ $\omega_t = \beta_0 + \beta_1 h_t$ $h_t = \text{SRN}(x_t, h_{t-1}) \text{ with } x_t = (\omega_{t-1}, e^{y_{t-1}}, \sigma_{t-1}^2)$	$0 \leq \beta < 1$

Table 1: Several specifications of RECH models.

Then,

$$\mathbb{V}(y_t | \sigma_0^2) \leq \frac{M}{1 - \alpha - \beta} + \sigma_0^2, \text{ for all } t \geq 0.$$

That is, if the initial volatility σ_0^2 is finite almost surely, then all the subsequent y_t have a finite variance almost surely.

The proof can be found in the Appendix. The conditions in the theorem prevent the volatility σ_t^2 from exploding and not cause numerical issues in training. The condition that $\alpha > 0, \beta > 0, \alpha + \beta < 1$ is standard in the GARCH literature and easy to impose. The second condition, i.e. the finite recurrent component condition, is imposed by using the bounded ReLU activation function (to bound h_t), and assuming a bounded support for β_0 and β_1 (we used uniform $U(0,0.5)$ for these two parameters).

4 Bayesian inference

This section discusses Bayesian estimation and inference for RECH models. We are interested in sampling from the posterior distribution

$$\pi(\theta) = p(\theta | y_{1:T}) = \frac{p(y_{1:T} | \theta)p(\theta)}{p(y_{1:T})}, \quad (9)$$

where $p(y_{1:T} | \theta)$ is the likelihood function, $p(\theta)$ is the prior and $p(y_{1:T}) = \int_{\Theta} p(y_{1:T} | \theta)p(\theta)d\theta$ is the marginal likelihood. Recall that the vector of model parameters θ consists of the recurrent and GARCH parameters. For example, the SRN-GARCH specification in (8a)-(8d) has the nine parameters $\theta = (\beta_0, \beta_1, \alpha, \beta, v_0, v_1, v_2, w, b)$.

4.1 Sequential Monte Carlo (SMC)

The SMC method is an attractive approach for Bayesian inference and forecasting in volatility modelling (Li et al., 2020). SMC can sample efficiently from non-standard posteriors, provides the marginal likelihood estimate as a by-product, and is a convenient way for computing one-step-ahead forecasts. In order to sample from the posterior $\pi(\theta)$, the SMC method (Neal, 2001;

Del Moral et al., 2006; Chopin, 2002) first samples a set of M weighted particles $\{W_0^j, \theta_0^j\}_{j=1}^M$ from an easy-to-sample distribution $\pi_0(\theta)$, such as the prior $p(\theta)$, and then traverses these particles through intermediate distributions $\pi_t(\theta)$, $t = 1, \dots, K$, which become the posterior distribution $\pi(\theta)$, i.e. $\pi_K(\theta) = \pi(\theta)$. In our article, we set $\pi_0(\theta) = p(\theta)$ as the prior $p(\theta)$ if it is possible to sample from $p(\theta)$. There are two common ways to design such a sequence of intermediate distributions: likelihood annealing (Neal, 2001) and data annealing (Chopin, 2002). The SMC with likelihood annealing uses the following intermediate distributions

$$\pi_t(\theta) := \pi_t(\theta|y_{1:T}) \propto p(y_{1:T}|\theta)^{\gamma_t} p(\theta), \quad (10)$$

where γ_t is referred to as the temperature level and $0 = \gamma_0 < \gamma_1 < \gamma_2 < \dots < \gamma_K = 1$.

The SMC method consists of three main steps: reweighting, resampling and a Markov move. There are various ways to implement SMC in practice; here we briefly present one of these. At the beginning of iteration t , the set of weighted particles $\{W_{t-1}^j, \theta_{t-1}^j\}_{j=1}^M$ that approximate the intermediate distribution $\pi_{t-1}(\theta)$ is reweighted to approximate the target $\pi_t(\theta)$. The efficiency of these weighted particles as a representation of $\pi_t(\theta)$ is often measured by the effective sample size (ESS) (Kass et al., 1998; Liu and Chen, 1998) defined in (13). If the ESS is below a prespecified threshold, the particles are resampled; the resulting equally-weighted samples are then refreshed by a Markov kernel whose invariant distribution is $\pi_t(\theta)$. Algorithm 1 summarizes this SMC using the likelihood annealing method. We follow Gunawan et al. (2020) and choose the tempering sequence γ_t adaptively to ensure a sufficient level of particle efficiency by selecting the next value of γ_t such that ESS stays above a threshold.

SMC with likelihood annealing sampler is suitable for in-sample analysis, as it uses the sequence of distributions in (10) which requires the full training data $y_{1:T}$ to be available. For out-of-sample rolling forecasts where the model parameters θ are updated once new data arrive, it is necessary to use SMC with the data annealing (Chopin, 2002). This SMC sampler generates weighted particles from the following sequence of distributions

$$\pi_t(\theta) := \pi_t(\theta|y_{1:t}) \propto p(y_{1:t}|\theta)p(\theta) \propto \pi_{t-1}(\theta)p(y_t|\theta, y_{1:t-1}), \quad (16)$$

with $y_{1:t}$ the data available up to time t . The unnormalized weights at the SMC step t in (11) become

$$w_t^j = W_{t-1}^j \frac{p(y_{1:t}|\theta_{t-1}^j)p(\theta_{t-1}^j)}{p(y_{1:t-1}|\theta_{t-1}^j)p(\theta_{t-1}^j)} = W_{t-1}^j p(y_t|y_{1:t-1}, \theta_{t-1}^j), \quad j = 1, \dots, M. \quad (17)$$

Algorithm 2 in the Appendix summarizes SMC with data annealing. For RECH models, we use SMC with likelihood annealing for in-sample Bayesian analysis, and SMC with data annealing for out-of-sample analysis and forecasting.

4.2 Model choice by marginal likelihood

The marginal likelihood is often used to choose between models using the Bayes factor (Jeffreys, 1935; Kass and Raftery, 1995). In order to compare the relative performance between two models M_1 and M_2 on data $y_{1:T}$, we can use the Bayes factor

$$BF_{M_1, M_2} = \frac{p(y_{1:T}|M_1)}{p(y_{1:T}|M_2)}. \quad (18)$$

Algorithm 1 SMC with likelihood annealing for RECH models

1. Sample $\theta_0^j \sim p(\theta)$ and set $W_0^j = 1/M$ for $j = 1 \dots M$
2. **For** $t = 1, \dots, K$,

Step 1: Resampling: Compute the unnormalized weights

$$w_t^j = W_{t-1}^j \frac{p(y_{1:T} | \theta_{t-1}^j)^{\gamma_t} p(\theta_{t-1}^j)}{p(y_{1:T} | \theta_{t-1}^j)^{\gamma_{t-1}} p(\theta_{t-1}^j)} = W_{t-1}^j p(y_{1:T} | \theta_{t-1}^j)^{\gamma_t - \gamma_{t-1}}, \quad j = 1, \dots, M \quad (11)$$

and set the new normalized weights

$$W_t^j = \frac{w_t^j}{\sum_{s=1}^M w_t^s}, \quad j = 1, \dots, M. \quad (12)$$

Step 2: Compute the effective sample size (ESS)

$$\text{ESS} = \frac{1}{\sum_{j=1}^M (W_t^j)^2}. \quad (13)$$

if $\text{ESS} < cM$ for some $0 < c < 1$, **then**

- (i) **Resampling:** Resample from $\{\theta_{t-1}^j\}_{j=1}^M$ using the weights $\{W_t^j\}_{j=1}^M$, and then set $W_t^j = 1/M$ for $j = 1 \dots M$, to obtain the new equally-weighted particles $\{\theta_t^j, W_t^j\}_{j=1}^M$.
- (ii) **Markov move:** For each $j = 1, \dots, M$, move the sample θ_t^j according to N_{lkh} random walk Metropolis-Hastings steps:
 - (a) Generate a proposal $\theta_t^{j'}$ from a multivariate normal distribution $\mathcal{N}(\theta_t^j, \Sigma_t)$ with Σ_t the covariance matrix.
 - (b) Set $\theta_t^j = \theta_t^{j'}$ with the probability

$$\min \left(1, \frac{p(y_{1:T} | \theta_t^{j'})^{\gamma_t} p(\theta_t^{j'})}{p(y_{1:T} | \theta_t^j)^{\gamma_t} p(\theta_t^j)} \right); \quad (14)$$

otherwise keep θ_t^j unchanged.

end

3. The log of the estimated marginal likelihood is

$$\widehat{\log p(y_{1:T})} = \sum_{t=1}^K \log \left(\sum_{j=1}^M w_t^j \right). \quad (15)$$

The larger the Bayes factor BF_{M_1, M_2} , the stronger evidence that M_1 is more strongly supported by the data than M_2 . We note that the SMC with likelihood annealing sampler in the previous section provides an efficient way to compute the marginal likelihood.

4.3 Runtime of the SMC with likelihood annealing sampler

As SMC is parallelizable, the running time depends on how the algorithm is parallelized. For example, we can run the algorithm on a single multi-core machine or multiple multi-core cluster machines. Table 2 shows the runtime of the SMC sampler, with $M = 1000$ and $M = 10000$ particles, when sampling the GARCH and SRN-GARCH models using only one core and six cores (numbers in parentheses). We run all examples on a standard laptop with moderate specification: Intel Core i7, 16GB RAM, 2.2GHz and 6 cores. We use $M = 1000$ in this paper as this value is sufficient to obtain consistent estimation results.

For comparison, the table also shows the running time of the MCMC sampler for the GARCH model by the R package `bayesGARCH` of Ardia and Hoogerheide (2010) with two different numbers of iterations $N = 20,000$ and $N = 200,000$, and the runtime of the MCMC sampler for the SV model using the R package `stochvol` of Hosszejni and Kastner (2021) with $N = 50,000$ and $N = 500,000$. These values of M and N are selected such that the standard errors of the SMC estimators (characterized by the effective sample size) are similar to that of the MCMC estimators (measured by the Integrated Autocorrelation Time, IACT). Here, we use the CODA R package of Plummer et al. (2006) to compute the IACT of the MCMC chains obtained from the `bayesGARCH` and `stochvol` packages.

	SMC		bayesGARCH		stochvol	
	$M = 1000$	$M = 10000$	$N = 20000$	$N = 200000$	$N = 50000$	$N = 500000$
SRN-GARCH	150 (73)	1568 (475)				
GARCH	24 (34)	154 (400)	38	405		
SV					133	1398

Table 2: The running time (in seconds) for the GARCH, SRN-GARCH and SV models for analysing the SP500 dataset, using various sampling methods. For the SMC sampler, the numbers in parentheses show the corresponding runtime when using all 6 cores.

It is important to note that, unlike the MCMC sampler, SMC is parallelizable and hence its runtime can be reduced significantly when running on a computing cluster. As shown in Table 2, the SMC sampler for SRN-GARCH is quite computationally expensive when using a single CPU core. However, its runtime is significantly reduced when running in parallel with six cores. In all of our examples in Section 5, SMC was run on a high performance computing cluster.

5 Simulation study and applications

This section evaluates the in-sample and out-of-sample performance of RECH models on simulation and stock return datasets. We use the SMC with likelihood annealing to perform in-sample Bayesian inference and the SMC with data annealing to obtain one-step-ahead forecasts. Table 3 lists our implementation settings of the SMC samplers.

Variable	Description	Value
K	Number of annealing levels	10000
M	Number of particles	10000
c	Constant of the ESS threshold	0.8
N_{lik}	Number of Markov moves in the SMC with likelihood annealing	30
N_{data}	Number of Markov moves in the SMC with data annealing	30

Table 3: Implementation settings of the SMC samplers.

We now discuss the selection of the prior distributions for RECH models. We use a normal prior with a zero mean and variance 0.1 for the recurrence parameters (v_0, v_1, v_2, w, b) ; as empirical results from the deep learning literature show that the values of the weights of neural networks are often small. As a linear activation function for $g(\cdot)$ is used, we put the uniform prior $U(0,0.5)$ on β_0 and β_1 to impose the positivity of ω_t . For the GARCH parameters, we choose the same priors as we use for the corresponding GARCH-type models, as summarized in Table 4.

GARCH(1,1)		GJR(1,1)		EGARCH(1,1)	
Parameter	Prior	Parameter	Prior	Parameter	Prior
ω	$U(0,10)$	ω	$U(0,10)$	ω	$\mathcal{N}(0,1)$
α	$U(0,1)$	α	$U(0,1)$	α	$\mathcal{N}(0,1)$
β	$U(0,1)$	β	$U(0,1)$	β	$U(0,1)$
		γ	$\mathcal{N}(0,0.1)$	γ	$\mathcal{N}(0,0.1)$

Table 4: Prior distributions for the parameters in the GARCH(1,1), GJR(1,1) and EGARCH(1,1) models. The notations U and \mathcal{N} denote the Uniform and Gaussian distributions, respectively.

Next, we discuss the score metrics used to evaluate the out-of-sample performance. Denote by D_{test} a test dataset, T_{test} the number of observations in D_{test} and $\hat{\theta}$ the posterior mean estimate of θ ; we use four predictive scores to measure out of sample performance: the partial predictive score (PPS), the number of violations (#Vio.), the quantile score (QS) and the hit percentage (%Hit) to measure the out-of-sample performance. The PPS (Gneiting and Raftery, 2007) is evaluated on the test dataset D_{test} as

$$\text{PPS} := -\frac{1}{T_{\text{test}}} \sum_{D_{\text{test}}} \log p(y_t | y_{1:t-1}, \hat{\theta}).$$

The model with smallest PPS is preferred. The #Vio. is defined as the number of times over the test data D_{test} that the observation y_t is outside its 99% one-step-ahead forecast interval. One of the main applications of volatility modelling is to forecast the Value at Risk (VaR). The α -VaR is defined as the α -quantile of the one-step-ahead forecast distribution $p(y_t|y_{1:t-1}, \hat{\theta})$. The performance of a method producing VaR forecasts is often measured by the quantile score (Taylor, 2019) defined as

$$QS := \frac{1}{T_{test}} \sum_{D_{test}} (\alpha - I_{y_t \leq q_{t,\alpha}})(y_t - q_{t,\alpha}),$$

where $q_{t,\alpha}$ is the α -VaR forecast of y_t , conditional on $y_{1:t-1}$. The smaller the quantile score, the better the VaR forecast. The %Hit (Taylor, 2019) is defined as the percentage of the y_t in the test data that is below its α -VaR forecast. The %Hit is expected to be close to α , if the model predicts well. We note that these predictive performance measures complement each other. For example, it is possible to make the number of violations small by increasing the forecast volatility, but the PPS and QS scores then increase. A volatility model minimizing all three predictive scores, and having a hit percentage close to α , is arguably the preferred one.

For the simulation data, given the true volatility σ_t , we follow Hansen and Lunde (2005) and use six additional predictive scores as summarized in Table 5.

Score	Definition
MSE ₁	$T_{test}^{-1} \sum_{D_{test}} (\sigma_t - \hat{\sigma}_t)^2$
MSE ₂	$T_{test}^{-1} \sum_{D_{test}} (\sigma_t^2 - \hat{\sigma}_t^2)^2$
MAE ₁	$T_{test}^{-1} \sum_{D_{test}} \sigma_t - \hat{\sigma}_t $
MAE ₂	$T_{test}^{-1} \sum_{D_{test}} \sigma_t^2 - \hat{\sigma}_t^2 $
QLIKE	$T_{test}^{-1} \sum_{D_{test}} (\log(\hat{\sigma}_t^2) + \sigma_t^2 \hat{\sigma}_t^{-2})$
R ² LOG	$T_{test}^{-1} \sum_{D_{test}} [\log(\sigma_t^2 \hat{\sigma}_t^{-2})]^2$

Table 5: Definition of the predictive scores to measure the out-of-sample performance on simulation and index data. Here, $\hat{\sigma}_t$ is an estimate of the volatility σ_t .

5.1 Simulation studies

5.1.1 Simulation study I (SIM I)

We generated a time series of 2000 observations from the GARCH(1,1) model

$$y_t = \sigma_t \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0,1), \quad t=1, \dots, T, \tag{19a}$$

$$\sigma_t^2 = 0.05 + 0.18 y_{t-1}^2 + 0.8 \sigma_{t-1}^2, \quad t=2, \dots, T, \quad \sigma_1^2 = 0.1. \tag{19b}$$

The first 1000 observations are used for model estimation and the last 1000 observations for out-of-sample analysis. Table 6 shows the posterior means and the posterior standard deviations of the GARCH(1,1) and SRN-GARCH model parameters, obtained from the SMC using likelihood annealing. Figure 5 plots the estimated volatility together with the true volatility of the simulated data, i.e. the true values σ_t^2 generated from equation (19b).

	ω	α	β	β_0	β_1	v_1	v_2	log ML
GARCH	0.048 (0.022)	0.178 (0.029)	0.791 (0.035)					-1507.2 (0.118)
SRN-GARCH		0.150 (0.028)	0.806 (0.029)	0.056 (0.022)	0.232 (0.152)	0.154 (0.314)	-0.248 (0.378)	-1507.3 (0.152)

Table 6: SIM I: Posterior means (in bold) of the GARCH and SRN-GARCH model parameters with the posterior standard deviations in brackets. The last column shows the natural logarithms of the estimated marginal likelihood with the Monte Carlo standard errors in brackets, averaged over 10 different runs of the likelihood annealing SMC.

The estimation results in Table 6 and the volatility plots in Figure 5 suggest some important implications. First, the GARCH parameters α and β of the SRN-GARCH model in Table 6 are close to those of the GARCH model and the volatility estimated from the RECH model in Figure 5 are close to the true volatility, implying that the estimated RECH model is close to the GARCH(1,1) model if the GARCH(1,1) is the true model. The almost identical estimates of the marginal likelihood in the last column of Table 6 also indicate that the GARCH(1,1) and RECH models provide an equally good fit to the data. Second, the coefficient β_1 is statistically insignificant, i.e. the posterior mean is less than two standard deviations from zero, and the values of the recurrent component ω_t are consistently small at all time points as Figure 5 shows, implying that there are no volatility effects other than linearity that are captured by the recurrent component ω_t and that the recurrent component ω_t contributes very little to the conditional variance at all time steps. Additionally, the weights v_1 and v_2 of the inputs of the recurrent component are statistically insignificant, suggesting that there is no evidence of non-linearity, long range dependence and leverage effects within the data generating process GARCH(1,1).

5.1.2 Simulation study II (SIM II)

We generated a time series of 2000 observations from the following non-linear GARCH-type model

$$y_t = \sigma_t \epsilon_t, \epsilon_t \sim \mathcal{N}(0,1), t=1,2,\dots,T, \quad (20a)$$

$$\sigma_t^2 = 0.05 + 0.10 y_{t-1}^2 + 0.21 \frac{y_{t-1}^2}{1+y_{t-1}^2} + 0.8 \sigma_{t-1}^2 + 0.11 \frac{\sigma_{t-1}^2}{1+\sigma_{t-1}^2} + 0.21 I_{[y_{t-1}<0]} y_{t-1}^2 + 0.1 \frac{I_{[y_{t-1}<0]}}{1+e^{-y_{t-1}^2}}. \quad (20b)$$

The model in (20a)-(20b) modifies the GJR(1,1) model by adding non-linear transformations of the past observation, conditional variance and leverage term to the equation of the

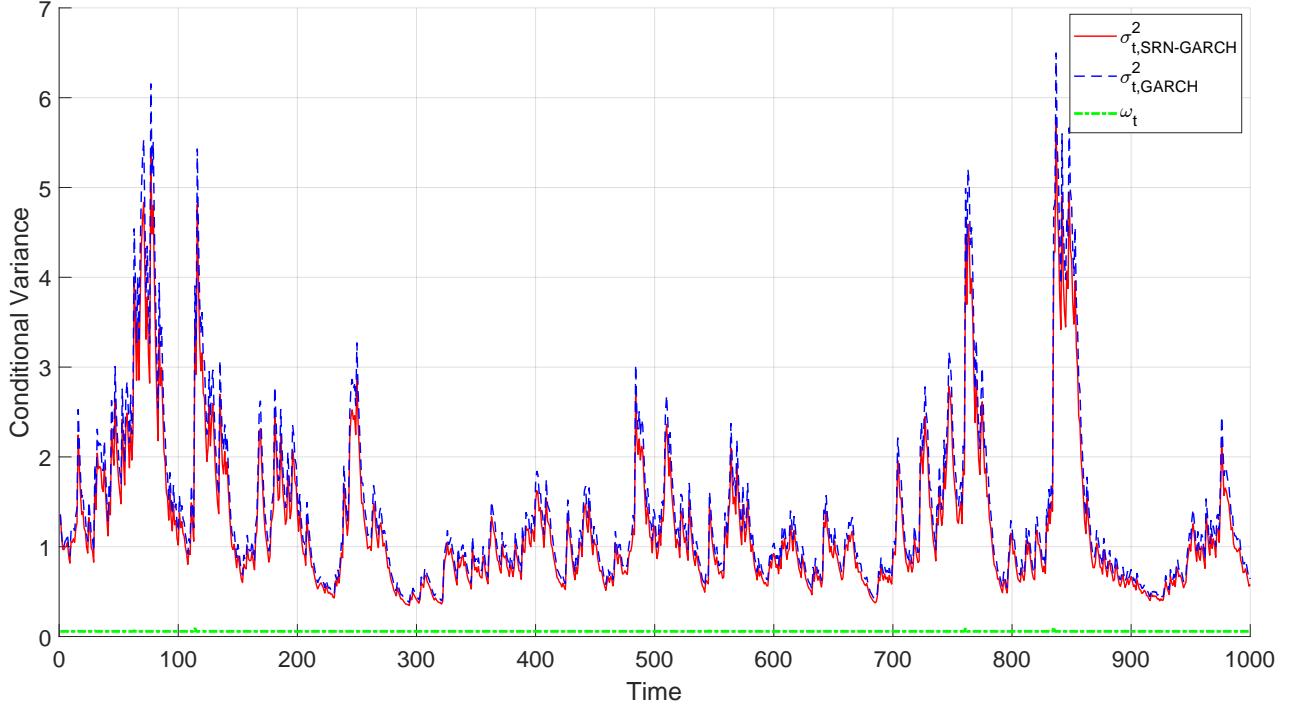


Figure 5: SIM I: The true conditional variance (dashed line) and the estimated conditional variance (solid line) using the SRN-GARCH model. The bottom line shows the values of the recurrent component ω_t of the SRN-GARCH specification at all time points. (The figure is better viewed in colour).

conditional variance. The volatility evolution in (20b) suggests that the simulated volatility exhibits highly non-linear effects. The parameters of the model in (20a)-(20b) are set so that the simulated data somewhat resembles real financial time series data exhibiting both volatility clustering and leverage effects. The first $T = 1000$ observations are used for model estimation and the last 1000 for out-of-sample analysis. Table 7 shows the posterior means and standard deviations of the parameters from the GARCH(1,1), GJR(1,1), EGARCH(1,1) and three RECH counterparts.

The estimation results from Table 7 suggest the following. First, the posterior means of α and β from the GARCH and GJR models are close to their true values, suggesting that the GARCH and GJR models can capture the linear serial dependence within the volatility dynamics of the data generating process. The constants ω of both the GARCH and GJR model are significantly inflated compared to the true value, possibly caused by the non-linear effects that cannot be captured by the GARCH and GJR models. The leverage parameter γ of the GJR model is close to zero and statistically insignificant, implying that the GJR model cannot capture the leverage effect in the data generating process. The leverage parameter γ of the EGARCH model is more than three standard deviations from zero, implying that the EGARCH model is the only benchmark GARCH-type model that can capture the simulated leverage effect.

Second, the coefficients β_1 of the SRN-GARCH and SRN-GJR models are more than three standard deviations from zero, implying that there is strong evidence of non-linearity in the volatility dynamics, and that the RNN structure within the recurrent component ω_t of the

	ω	α	β	γ	β_0	β_1	v_1	v_2	Log ML
GARCH	0.854 (0.124)	0.188 (0.018)	0.807 (0.018)						-3617.9 (0.164)
SRN-GARCH		0.246 (0.048)	0.556 (0.158)		0.328 (0.120)	0.382 (0.100)	-0.525 (0.253)	0.396 (0.232)	-3614.9* (0.232)
GJR	0.846 (0.130)	0.204 (0.035)	0.801 (0.021)	-0.012 (0.024)					-3620.1 (0.141)
SRN-GJR		0.141 (0.045)	0.557 (0.138)	0.179 (0.064)	0.354 (0.110)	0.373 (0.092)	-0.180 (0.294)	-0.418 (0.172)	-3611.6* (0.211)
EGARCH	0.106 (0.027)	0.373 (0.041)	0.975 (0.005)	-0.101 (0.026)					-3616.3 (0.147)
SRN-EGARCH	-0.058 (0.173)	0.450 (0.145)	0.976 (0.017)	-0.114 (0.037)	0.227 (0.140)	0.270 (0.139)	-0.028 (0.361)	0.211 (0.360)	-3613.1* (0.202)

Table 7: SIM II: Posterior means (in bold) of the GARCH and RECH model parameters with the posterior standard deviations in brackets. The last column shows the natural logarithms of the estimated marginal likelihood with the Monte Carlo standard errors in brackets, averaged over 10 different runs of the SMC with likelihood annealing sampler. The asterisks indicate when the Bayes factors strongly support RECH models over their GARCH-type counterparts.

SRN-GARCH and SRN-GJR model is able to capture such dependence. The weight v_1 with respect to the leverage input y_{t-1} of the RNN in the SRN-GARCH model is more than two standard deviations from zero and the leverage parameters γ of the SRN-GJR and SRN-EGARCH are more than three standard deviation from zero, all suggesting that the three RECH specifications can capture the leverage effects exhibited within the simulated volatility. The recurrent component ω_t of RECH models is useful in capturing the leverage effects overlooked by the GARCH and GJR models. Interestingly, the coefficient v_2 with respect to the input σ_{t-1}^2 in the SRN-GJR is more than two standard deviations from zero, indicating that the SRN-GJR is able to detect the non-linearity dependence of the past conditional variance on the current conditional variance σ_t^2 , exhibited within the simulated data generating process. The analysis above suggests that observing the recurrent parameters of RECH models, e.g. β_1, v_1 and v_2 , helps to detect the possible non-linearity effects within the underlying volatility.

Third, the estimates of marginal likelihood in the last column of Table 7 show that RECH models consistently have higher marginal likelihood than their GARCH-type counterparts and that the SRN-GJR model provides the best fit to the data. The difference between the log marginal likelihood estimates is equivalent to the Bayes factors of the SRN-GARCH, SRN-GJR and SRN-EGARCH models compared to the GARCH, GJR and EGARCH models of roughly $e^3 \approx 20.1$, $e^8 \approx 3000$ and e^3 , respectively, strongly supporting the RECH models. We note that among the benchmark GARCH-type models, the EGARCH model best fits to the SIM II data.

Figure 6 plots the values of the recurrent component and estimated volatility of the SRN-GJR model, together with the true volatility, at all time points. Figure 11 in the Appendix plots the volatility estimated by the GJR model and the true volatility. During the low volatility periods, i.e. time t is between 0-100 or 900-1000, the volatility of the SRN-GJR

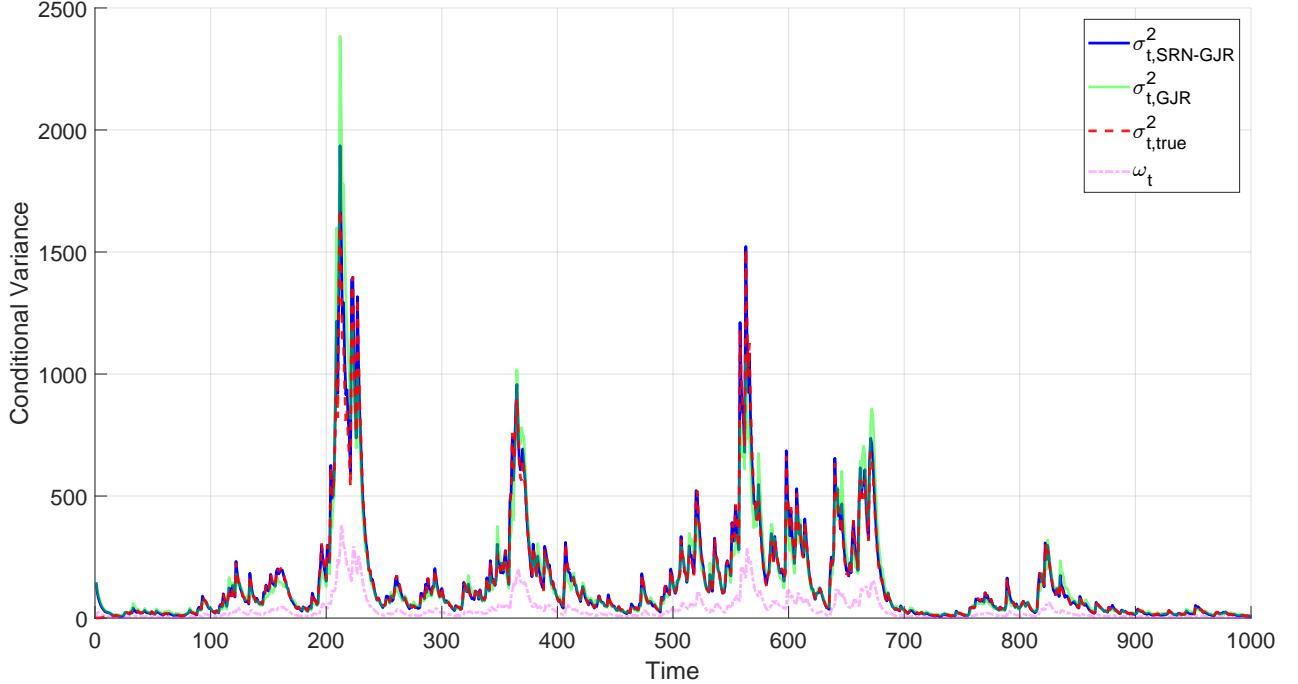


Figure 6: SIM II: The true conditional variance (dashed line) and the estimated conditional variance (solid lines) by the GJR and SRN-GJR models. The bottom line shows the values of the recurrent component ω_t of the SRN-GJR specification. The SRN-GJR plot appears to trace the true volatility plot better than the GJR plot. (The figure is better viewed in colour).

model are very close to the true volatility, which is also the case for the GJR model in Figure 11. During the periods when the volatility changed dramatically and oscillated highly, i.e. t is between 200-250, 350-400 or 500-700, the volatility produced by the SRN-GJR model still tracks the true volatility well, but this is not the case for the GJR model. The GJR model often produces overly-large and overly-small volatility during these periods of abruptly changed volatility. The SRN-GJR model, on the other hand, appears to be able to capture well these changes. The plot of the recurrent component ω_t at all time points in Figure 6 shows that it is highly responsive to the changes in the true volatility.

Table 8 reports the forecast performance of the benchmark GARCH-type and the RECH models. For the SIM I data, the forecast performance of the SRN-GARCH model is very close to that of the GARCH model, in all predictive scores, which again strongly supports the earlier in-sample conclusion that the SRN-GARCH model closely approximates the GARCH model if the GARCH is the true model. For the SIM II data, Table 8 suggests some important results. First, the RECH models outperform their GARCH-type counterparts in most of the predictive scores, which is consistent with the in-sample analysis showing that the RECH models fit better than their benchmark GARCH-type counterparts. Second, the SRN-GJR model has the best forecast performance for all predictive measures and the SRN-EGARCH model also performs well on the SIM II data, which is consistent with the in-sample analysis showing that these two RECH specifications have the highest marginal likelihood estimates. Third, amongst the benchmark GARCH-type models, the EGARCH model has the best predictive performance, compared to the GARCH and GJR models.

	PPS	#Vio	QS	%Hit	MSE ₁	MSE ₂	MAE ₁	MAE ₂	QLike	R ² Log
SIM I										
GARCH*	1.794	09	0.044	0.011	0.001	0.020	0.016	0.069	1.689	0.001
SRN-GARCH	1.795	09	0.044	0.011	0.001	0.057	0.022	0.101	1.690	0.001
SIM II										
GARCH	3.013	14	0.145	0.008	0.390	88.970	0.485	6.018	4.207	0.046
SRN-GARCH*	3.009	12	0.142	0.008	0.256	85.151	0.349	4.739	4.195	0.022
GJR	3.011	16	0.144	0.008	0.331	77.430	0.445	5.503	4.204	0.039
SRN-GJR*	3.003	10	0.139	0.009	0.023	5.071	0.114	1.339	4.186	0.003
EGARCH	3.005	10	0.142	0.006	0.109	22.250	0.251	3.078	4.191	0.013
SRN-EGARCH*	3.004	10	0.140	0.008	0.054	12.876	0.173	2.153	4.188	0.007

Table 8: Simulation: one-step-ahead forecast comparison. For the QS and %Hit measures, the results are calculated at the 1%-quantile. For the SIM II data, the bold numbers denote the best scores. For each pair of the RECH and GARCH-type models, the asterisk indicates the models having better forecast performance.

5.1.3 Simulation study III (SIM III)

This simulation study examines if RECH models are able to simulate the long-memory volatility, by fitting the FIGARCH(1,d,1) model of Baillie et al. (1996) to data generated from RECH models. The FIGARCH(1,d,1) model is defined as:

$$y_t = \sigma_t \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0,1), \quad t=1,2,\dots,T,$$

$$\sigma_t^2 = \omega + \beta \sigma_{t-1}^2 + [1 - \beta L - (1 - \psi L)(1 - L)^d] y_t^2,$$

where $d \in (0,1)$ is the fractional integrated parameter and L is the backshift operator. The parameter $\Theta = (\omega, \psi, d, \beta)$. When $d=0$, the FIGARCH becomes a GARCH model. When $d>0$ and is close to 1, the persistence of the past shocks in the FIGARCH process decays at a slow hyperbolic rate (Baillie et al., 1996); hence the FIGARCH process exhibits long-memory effects in its volatility dynamics.

	α	β	β_0	β_1	v_0	v_1	v_2	w	b
θ_1	0.058	0.681	0.068	0.418	-0.018	-0.430	0.524	0.161	-0.173
θ_2	0.071	0.690	0.075	0.362	0.062	-0.422	0.538	0.087	-0.130
θ_3	0.076	0.744	0.016	0.388	-0.075	-0.574	0.400	-0.040	-0.023
θ_4	0.057	0.562	0.101	0.413	0.015	-0.380	0.652	0.270	-0.170

Table 9: SIM III: The parameters used in the DGP.

We use the SRN-GARCH model as the true data generating process (DGP) with the four different parameter sets $\theta_i, i = 1,..,4$, listed in Table 9. These are the estimated parameters obtained in Section 5.2 when the SRN-GARCH model is fitted to four real datasets. For each parameter set θ_i , 500 datasets of $T=3000$ observations are generated from each of two different

specifications of the SRN-GARCH model: $\beta_1=0$ and β_1 equals to the true values in Table 9, i.e. $\beta_1\neq 0$. We note that if $\beta_1=0$, the DGP is the GARCH(1,1) model. To generate each time series, we generate 10,000 observations and use the last 3,000 for the simulation data. We then use the Matlab MFE toolbox¹, with the default settings, to produce the Quasi-Maximum Likelihood Estimate (QMLE) of the parameter $\widehat{\Theta}_i, i=1,\dots,4$, of the FIGARCH(1,d,1) model. Table 10 shows the means and standard deviations averaged over 500 QMLE estimates of the FIGARCH(1,d,1) parameters.

	DGP is GARCH(1,1) ($\beta_1=0$)				DGP is SRN-GARCH ($\beta_1\neq 0$)			
	$\widehat{\omega}$	$\widehat{\psi}$	\widehat{d}	$\widehat{\beta}$	$\widehat{\omega}$	$\widehat{\psi}$	\widehat{d}	$\widehat{\beta}$
$\widehat{\Theta}_1$	0.211 (0.043)	0.423 (0.159)	0.017 (0.039)	0.373 (0.127)	0.088 (0.014)	0.107 (0.052)	0.728 (0.109)	0.708 (0.078)
$\widehat{\Theta}_2$	0.225 (0.064)	0.366 (0.206)	0.036 (0.053)	0.323 (0.155)	0.126 (0.020)	0.152 (0.058)	0.608 (0.12)	0.647 (0.100)
$\widehat{\Theta}_3$	0.045 (0.018)	0.210 (0.214)	0.093 (0.059)	0.222 (0.162)	0.095 (0.016)	0.105 (0.053)	0.711 (0.110)	0.653 (0.089)
$\widehat{\Theta}_4$	0.235 (0.019)	0.437 (0.127)	0.002 (0.014)	0.376 (0.125)	0.096 (0.017)	0.124 (0.054)	0.706 (0.116)	0.711 (0.084)

Table 10: SIM III: Means and standard deviations (in brackets) of 500 QMLE estimates of the FIGARCH(1,d,1) model parameters.

The important conclusion from Table 10 is that the short-memory and long-memory properties of the GARCH(1,1) and SRN-GARCH models, respectively, are distinguishable. When the DGP is the GARCH(1,1), the estimates of the fractional integrated parameter \widehat{d} are insignificant in all cases, suggesting that there is no evidence of the long-memory effects in the volatility of the GARCH(1,1) model. When the DGP is the SRN-GARCH model, i.e. $\beta_1\neq 0$, the estimates of the fractional integrated parameter \widehat{d} are close to 1 in all cases, implying the existence of long-memory in the volatility dynamics of simulated time series, which are generated from SRN-GARCH models. The difference in the QMLE estimates of the parameter d between the two DGPs in Table 10 implies that the SRN-GARCH model is able to simulate long-memory volatility effects. We observe similar results for the SRN-GJR and SRN-EGARCH models.

5.2 Applications to stock market returns

We demonstrate the performance of RECH models using four stock index datasets: the Standard and Poor's 500 Index (SP500), the Japanese Nikkei 225 Index (N225), the Russell 2000 Index (RUT) and the German stock index (DAX). The datasets were downloaded from the Realized Library of The Oxford-Man Institute². We used the daily closing prices $\{P_t, t=1,\dots,T_P\}$

¹<https://github.com/bashtage/mfe-toolbox/>

²<https://realized.oxford-man.ox.ac.uk/>

and calculated the demeaned return process as

$$y_t = 100 \left(\log \frac{P_{t+1}}{P_t} - \frac{1}{T_P - 1} \sum_{i=1}^{T_P-1} \log \frac{P_{i+1}}{P_i} \right), \quad t = 1, 2, \dots, T_P - 1. \quad (21)$$

The length of the four return series is fixed to be $T = 4000$, with $T = T_P - 1$, and each series is divided into an in-sample period of the first $T_{\text{in}} = 2000$ observations and an out-of-sample period of the last $T_{\text{out}} = 2000$ observations. Table 11 summarizes the datasets.

	In-sample Period	Out-of-sample Period	T_{in}	T_{out}
SP500	27 Feb 2004 – 06 Feb 2012	06 Feb 2012 – 24 Jan 2020	2000	2000
N225	16 Sep 2003 – 16 Nov 2011	17 Nov 2011 – 24 Jan 2020	2000	2000
RUT	24 Feb 2004 – 01 Feb 2012	02 Feb 2012 – 24 Jan 2020	2000	2000
DAX	06 Aug 2003 – 02 Nov 2011	02 Nov 2011 – 24 Jan 2020	2000	2000

Table 11: Description of the four index datasets.

Table 12 reports some descriptive statistics for these four datasets together with the modified R/S test (Lo, 1991) for long-range memory in the logarithm of the squared returns. Lo's modified R/S test is widely used in the financial time series literature; see, e.g., Lo (1991), Giraitis et al. (2003), Breidt et al. (1998). All the index data exhibit some negative skewness, a high excess kurtosis and high variation. The N225 returns are more skewed and leptokurtic than those of the SP500, RUT and DAX data. The result of Lo's modified R/S test for long-memory dependence with several different lags q indicates that there is significant evidence of long-memory dependence in the SP500, RUT and DAX stock indices. For the N225 data, however, the evidence of long memory is less clear as the null hypothesis of short memory for the squared returns is not rejected at the 5% level of significance when $q = 20$ and $q = 30$.

	Min	Max	Std	Skew	Kurtosis	$V_n(10)$	$V_n(20)$	$V_n(30)$
SP500	-9.351	10.220	1.307	-0.256	12.502	3.188*	2.412*	2.047*
						2.664*	2.040*	1.748*
N225	-10.563	11.658	1.224	-0.585	18.171	2.768*	2.171*	1.905*
						1.956*	1.566	1.415
RUT	-8.391	8.056	1.364	-0.130	8.764	3.065*	2.385*	2.055*
						2.459*	1.943*	1.691
DAX	-7.437	9.993	1.267	0.115	10.960	3.226*	2.501*	2.146*
						2.456*	1.926*	1.670

Table 12: Descriptive statistics for the demeaned returns of the SP500, N225, RUT and DAX datasets. $V_n(q)$, $q = 10, 20$ and 30 , are the test statistics of Lo's modified R/S test of long memory with lag q . Upper and lower values of the 3 last columns are Lo's test statistics for absolute and squared returns, respectively. The asterisks indicate significance at the 5% level.

The Realized Library provides different realized measures³ that can be used in financial

³See <https://realized.oxford-man.ox.ac.uk/documentation/estimators> for the list of the available realized measures

econometrics as a proxy to the latent σ_t^2 . We use the following six common realized measures including Realized Variance (RV) (Andersen and Bollerslev, 1998), Bipower Variation (BV) (Barndorff-Nielsen and Shephard, 2004), Median Realized Volatility (MedRV) (Andersen et al., 2012), Realized Kernel Variance (Barndorff-Nielsen et al., 2008) with the Non-Flat Parzen kernel (RKV_1), the Tukey-Hanning kernel (RKV_2) and the Two-Scale/Bartlett kernel (RKV_3), to evaluate the forecast performance of the volatility models using the predictive scores in Table 5. Shephard and Sheppard (2010) give more details about the Realized Library.

Denote by RV_t the realized measure of σ_t^2 at time t . As the realized measures ignore the variation of the prices overnight and sometimes the variation in the first few minutes of the trading day when recorded prices may contain large errors (Shephard and Sheppard, 2010), we follow Hansen and Lunde (2005) to scale the realized measure RV_t as

$$\tilde{\sigma}_t^2 = \hat{c} \cdot RV_t \text{ where } \hat{c} = \frac{T_{\text{out}}^{-1} \sum_{t=T_{\text{in}}+1}^T [y_t - E(y_t | \mathcal{F}_{t-1})]^2}{T_{\text{out}}^{-1} \sum_{t=T_{\text{in}}+1}^T RV_t}, \quad t = T_{\text{in}} + 1, 2, \dots, T, \quad (22)$$

with $E(y_t | \mathcal{F}_{t-1})=0$, and use $\tilde{\sigma}_t^2$ as the estimate of the latent conditional variance σ_t^2 ; see Table 11 for a definition of T_{in} and T_{out} used in our datasets. See Martens (2002) and Fleming et al. (2003) for a similar scaling estimator of the daily volatility.

5.2.1 In-sample analysis

Table 13 and 14 summarize the estimation results of fitting the benchmark GARCH-type models and their RECH counterparts to the SP500, N225, RUT and DAX datasets. The posterior mean estimates and posterior standard deviation estimates are obtained using the SMC with likelihood annealing sampler. We draw the following conclusions from the estimation results.

	ω	α	β	γ	β_0	β_1	v_1	v_2	Mar.llh
SP500									
GARCH	0.016 (0.004)	0.093 (0.011)	0.894 (0.012)						-2778.3 (0.113)
SRN-GARCH		0.057 (0.011)	0.562 (0.082)		0.101 (0.026)	0.413 (0.063)	-0.380 (0.076)	0.652 (0.167)	-2742.3* (0.284)
GJR	0.024 (0.004)	0.040 (0.011)	0.891 (0.009)	0.065 (0.011)					-2764.4 (0.121)
SRN-GJR		0.011 (0.008)	0.685 (0.109)	0.109 (0.026)	0.078 (0.036)	0.349 (0.102)	-0.222 (0.093)	0.581 (0.196)	-2736.6* (0.323)
EGARCH	0.004 (0.002)	0.136 (0.016)	0.978 (0.003)	-0.126 (0.013)					-2754.8 (0.133)
SRN-EGARCH	-0.196 (0.092)	0.106 (0.030)	0.972 (0.016)	-0.236 (0.044)	0.066 (0.027)	0.343 (0.099)	0.104 (0.103)	0.538 (0.198)	-2744.0* (0.225)
N225									
GARCH	0.033 (0.008)	0.138 (0.016)	0.839 (0.018)						-2800.1 (0.119)
SRN-GARCH		0.076 (0.016)	0.744 (0.051)		0.016 (0.013)	0.388 (0.076)	-0.574 (0.156)	0.400 (0.147)	-2766.8* (0.279)
GJR	0.048 (0.009)	0.060 (0.015)	0.835 (0.020)	0.100 (0.018)					-2787.5 (0.121)
SRN-GJR		0.066 (0.021)	0.734 (0.061)	0.029 (0.036)	0.027 (0.015)	0.386 (0.081)	-0.531 (0.146)	0.429 (0.183)	-2769.2* (0.313)
EGARCH	-0.003 (0.004)	0.215 (0.022)	0.967 (0.006)	-0.134 (0.015)					-2772.0 (0.133)
SRN-EGARCH	-0.088 (0.048)	0.255 (0.046)	0.990 (0.009)	-0.164 (0.019)	0.033 (0.020)	0.344 (0.117)	-0.355 (0.148)	0.434 (0.325)	-2772.5 (0.225)

Table 13: SP500 and N225 data: Posterior means (in bold) of the parameters with the posterior standard deviations (in brackets). The last column shows the estimated log marginal likelihood with the Monte Carlo standard errors in brackets, averaged over 10 different runs of the SMC using the likelihood annealing algorithm. The asterisks indicate the cases when the Bayes factors strongly support the RECH models over their corresponding GARCH-type models.

	ω	α	β	γ	β_0	β_1	v_1	v_2	Mar.llh
RUT									
GARCH	0.036 (0.009)	0.112 (0.015)	0.863 (0.019)						-3037.6 (0.117)
SRN-GARCH		0.071 (0.017)	0.690 (0.078)		0.075 (0.040)	0.362 (0.082)	-0.422 (0.118)	0.538 (0.203)	-3014.3* (0.284)
GJR	0.049 (0.009)	0.048 (0.013)	0.864 (0.017)	0.084 (0.015)					-3025.4 (0.121)
SRN-GJR		0.019 (0.008)	0.780 (0.109)	0.109 (0.026)	0.078 (0.036)	0.311 (0.102)	-0.255 (0.093)	0.428 (0.196)	-3010.6* (0.323)
EGARCH	0.013 (0.004)	0.162 (0.021)	0.969 (0.006)	-0.111 (0.015)					-3023.1 (0.130)
SRN-EGARCH	-0.090 (0.049)	0.156 (0.027)	0.990 (0.010)	-0.161 (0.019)	0.050 (0.018)	0.307 (0.018)	-0.282 (0.205)	0.238 (0.254)	-3015.5* (0.225)
DAX									
GARCH	0.019 (0.005)	0.096 (0.013)	0.890 (0.014)						-2902.0 (0.119)
SRN-GARCH		0.058 (0.019)	0.681 (0.126)		0.068 (0.037)	0.418 (0.059)	-0.430 (0.131)	0.524 (0.281)	-2867.2* (0.279)
GJR	0.031 (0.006)	0.046 (0.010)	0.884 (0.013)	0.066 (0.013)					-2889.1 (0.120)
SRN-GJR		0.035 (0.016)	0.711 (0.088)	0.072 (0.032)	0.067 (0.038)	0.369 (0.095)	-0.301 (0.117)	0.575 (0.213)	-2866.4* (0.313)
EGARCH	0.006 (0.002)	0.158 (0.018)	0.975 (0.004)	-0.116 (0.014)					-2872.5 (0.128)
SRN-EGARCH	-0.063 (0.063)	0.169 (0.026)	0.989 (0.011)	-0.127 (0.020)	0.034 (0.024)	0.265 (0.132)	-0.185 (0.160)	0.193 (0.394)	-2869.0* (0.232)

Table 14: RUT and DAX data: Posterior means (in bold) of the parameters with the posterior standard deviations (in brackets). The last column shows the estimated log marginal likelihood with the Monte Carlo standard errors in brackets, averaged over 10 different runs of the SMC using the likelihood annealing algorithm. The asterisks indicate the cases when the Bayes factors strongly support the RECH models over their corresponding GARCH-type models.

First, the marginal likelihood estimates show that the RECH models fit the index datasets better than the GARCH-type models, except for the SRN-EGARCH model for the N225 data. For example, for the SP500 data, the Bayes factors of the SRN-GARCH, SRN-GJR and SRN-EGARCH models compared to the GARCH, GJR and EGARCH models are roughly e^{36} , e^{28} and e^{10} , respectively, which, according to Jeffrey's scale for interpreting the Bayes factor (Jeffreys, 1961), decisively support the RECH models. Among the benchmark GARCH-type models, the EGARCH model constantly has the highest marginal likelihood.

Second, the estimated posterior means of the parameter β_1 of the RECH models are

more than two standard deviations from zero in all cases, providing evidence of the volatility effects rather than linearity, e.g. probably non-linearity and long-memory effects, in the volatility dynamics and also suggesting that the recurrent component of the RECH models is able to effectively detect these effects. Additionally, the coefficients v_2 of the RECH models are statistically significant, indicating that the RECH models are able to detect the serial dependence rather than linearity that the previous conditional variance σ_{t-1}^2 has on σ_t^2 .

Third, the existence of the leverage effects in the volatility is clear across all four stock markets. The leverage parameters γ of the GJR and EGARCH models are statistically significant, implying that these models can detect the asymmetric volatility. All the leverage effect-related parameters γ and v_1 in the RECH models are statistically significant, except the parameter v_1 of the SRN-EGARCH model. In particular, the *linear* leverage coefficient γ of the SRN-GRJ and SRN-EGARCH models are significant, similarly to those of the GRJ and EGARCH models. Interestingly, the *non-linear* leverage coefficient v_1 of the RECH models is significant in almost all cases, suggesting that the spillover effect of asymmetric volatility can be non-linear. In particular, the leverage coefficient v_1 of the SRN-GARCH model is also statistically significant across all markets; i.e., unlike the conventional GARCH model, SRN-GARCH can detect the leverage effects in volatility.

Finally, as pointed out by a reviewer, in the EGARCH case the α and β appear to be less affected by adding the recurrent component. This is because the EGARCH structure is very different from GARCH and GJR (see Table 1); the EGARCH part in SRN-EGARCH is already non-linear in σ_{t-1} and hence it can accommodate non-linearity. This doesn't mean that SRN-EGARCH cannot improve EGARCH - β_1 is still far away from zero and the improvement is evident in better marginal likelihood and better out-of-sample prediction as confirmed in Section 5.2.2.

Figure 7 shows the volatility estimated by the GARCH and SRN-GARCH models for the SP500 index data, together with the values of the recurrent component ω_t at all time points. Figure 12 and 13 in the Appendix are similar plots for the SRN-GJR and SRN-EGARCH models. Clearly, the recurrent component ω_t is responsive to the changes in the volatility dynamics: it is small during the low volatility periods and large in the high volatility periods. This distinct behavior of financial volatility is well-captured by the recurrent neural network structure of the recurrent component ω_t .

Figure 8 plots the standardized residuals $\hat{\epsilon}_t$ from the GARCH and SRN-GARCH models together with their QQ-plots.

We observe similar results for the SRN-GRJ and GRJ models, SRN-EGARCH and EGARCH models. Generally the RECH residuals appear to lie closer to the expected straight line than those of the counterpart GARCH-type models.

Table 15 provides the skewness and kurtosis statistics together with the p -values of the Ljung-Box (LB) autocorrelation test of the residuals and squared residuals estimated by the RECH and the benchmark GARCH-type models. The p -value of the LB test, together with the sample ACF plots, of the standardized and squared standardized residuals suggest that there is no evidence of autocorrelation. The residuals produced by all models in Table 15 exhibit some negative skewness and have kurtosis values higher than 3 (the kurtosis of the standard normal distribution). In general, the residuals of the RECH models seem closer to normality than those of the corresponding GARCH-type models. Similarly to the GARCH-type models, it is straightforward to use Student's t distribution for the innovation in the

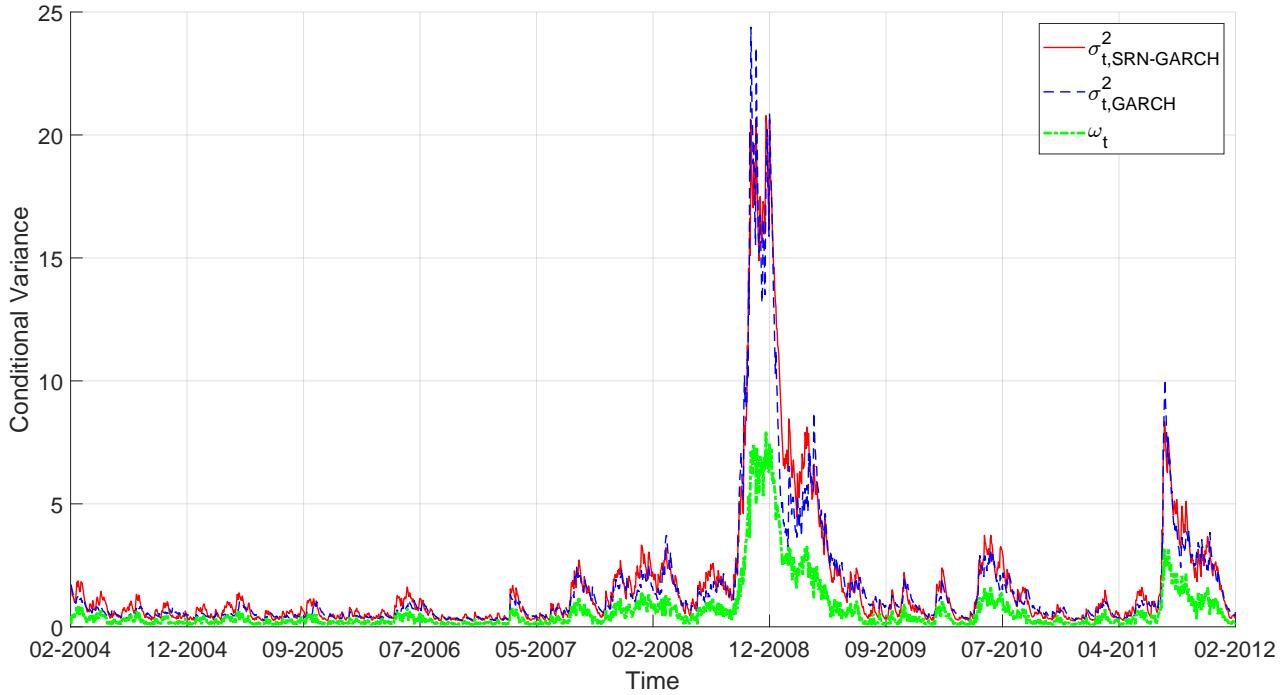


Figure 7: SP500: The in-sample conditional variance of the GARCH (dashed line) and SRN-GARCH (solid line) at all time points. The bottom line shows the values of the recurrent component ω_t of the SRN-GARCH specification. (The figure is better viewed in colour).

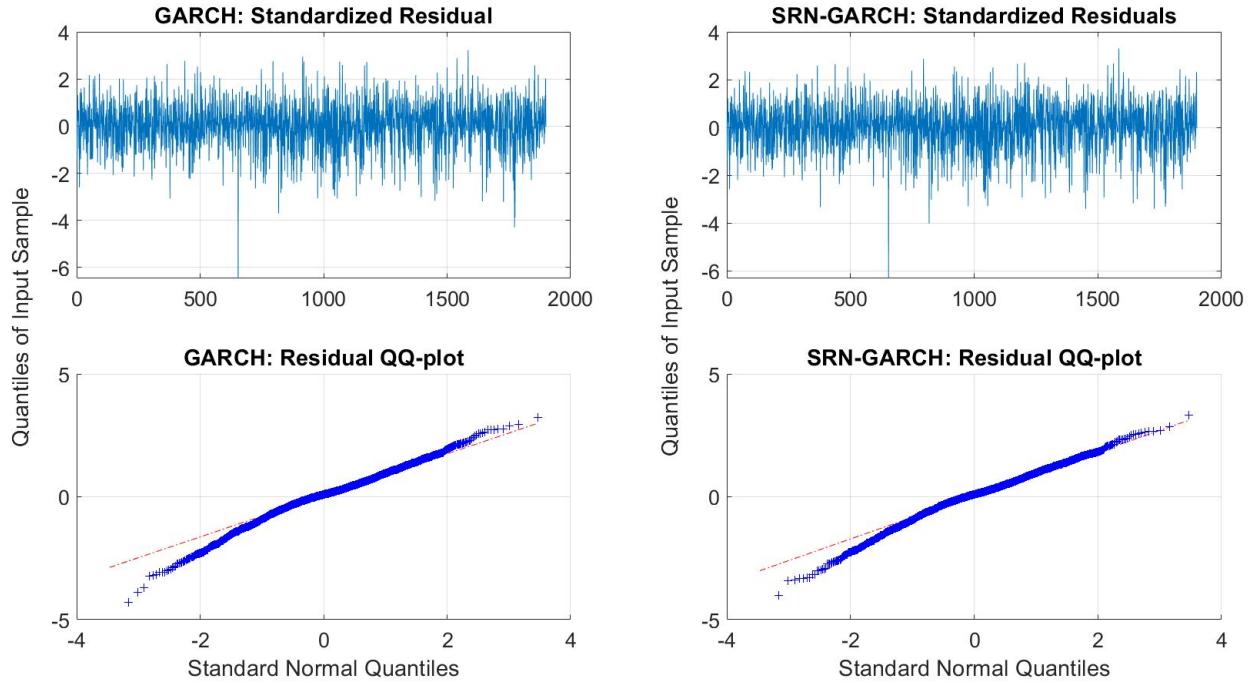


Figure 8: SP500: Estimated residuals $\hat{\epsilon}_t$ of the SRN-GARCH and GARCH models and their Q-Q plots.

	Fitted Conditional Variance				Residual $\hat{\epsilon}_t$			
	Mean	Std	Skew	Kurtosis	Std	Skew	Kurtosis	LB- $\hat{\epsilon}_t$
S&P500								
GARCH	1.650	2.845	4.543	26.522	1.002	-0.505	4.535	0.054
SRN-GARCH	1.755	2.910	4.045	21.251	1.002	-0.509	4.219	0.119
GJR	1.423	2.342	4.585	27.169	1.001	-0.485	4.202	0.058
SRN-GJR	1.600	2.877	4.445	25.585	1.001	-0.515	4.200	0.110
EGARCH	1.534	2.318	4.291	25.389	0.999	-0.587	4.616	0.051
SRN-EGARCH	1.603	2.887	4.661	29.080	0.999	-0.524	4.125	0.107
N225								
GARCH	1.493	3.133	7.792	72.005	0.999	-0.528	5.037	0.138
SRN-GARCH	1.390	2.521	6.950	58.482	0.999	-0.407	4.275	0.108
GJR	1.352	2.692	7.780	72.782	1.003	-0.455	4.509	0.104
SRN-GJR	1.434	2.634	7.054	60.510	1.003	-0.404	4.235	0.085
EGARCH	1.370	2.282	7.252	68.428	1.001	-0.404	4.197	0.073
SRN-EGARCH	1.242	2.430	7.605	73.471	1.001	-0.378	4.158	0.083
RUT								
GARCH	1.803	2.524	4.177	23.108	1.000	-0.318	3.717	0.046
SRN-GARCH	1.826	2.407	3.746	18.897	1.000	-0.379	3.728	0.084
GJR	1.619	2.122	4.343	25.208	1.002	-0.322	3.663	0.034
SRN-GJR	1.831	2.635	4.000	21.316	1.002	-0.364	3.725	0.085
EGARCH	1.681	1.973	3.775	20.933	0.999	-0.397	3.843	0.061
SRN-EGARCH	1.634	2.246	4.245	24.624	0.999	-0.373	3.690	0.057
DAX								
GARCH	1.598	2.203	4.332	26.122	1.002	-0.462	4.870	0.957
SRN-GARCH	1.656	2.378	3.484	16.904	1.002	-0.468	4.323	0.893
GJR	1.400	1.739	3.815	19.772	1.001	-0.399	4.478	0.957
SRN-GJR	1.510	2.013	3.517	17.005	1.001	-0.420	4.228	0.911
EGARCH	1.504	1.766	3.438	17.766	1.000	-0.414	4.116	0.905
SRN-EGARCH	1.262	1.730	4.132	23.497	1.000	-0.385	4.066	0.916

Table 15: SP500: Model diagnostics of the fitted conditional variance and residual $\hat{\epsilon}_t$. The LB p-values denote the p-value from the Ljung-Box test with 10 lags.

RECH models to improve the residual diagnostics; however, this extension is not considered here.

5.2.2 Out-of-sample analysis

Figure 9 plots the one-step-ahead forecast conditional variance of the GARCH and SRN-GARCH models, together with the out-of-sample realized variance of the S&P500 data, obtained by data annealing SMC. Figure 14 and 15 in the Appendix are similar plots for the case of the SRN-GJR and SRN-EGARCH models, respectively. To save space, we do not report the plots for the N225, RUT and DAX datasets as similar behaviors of the forecast variance

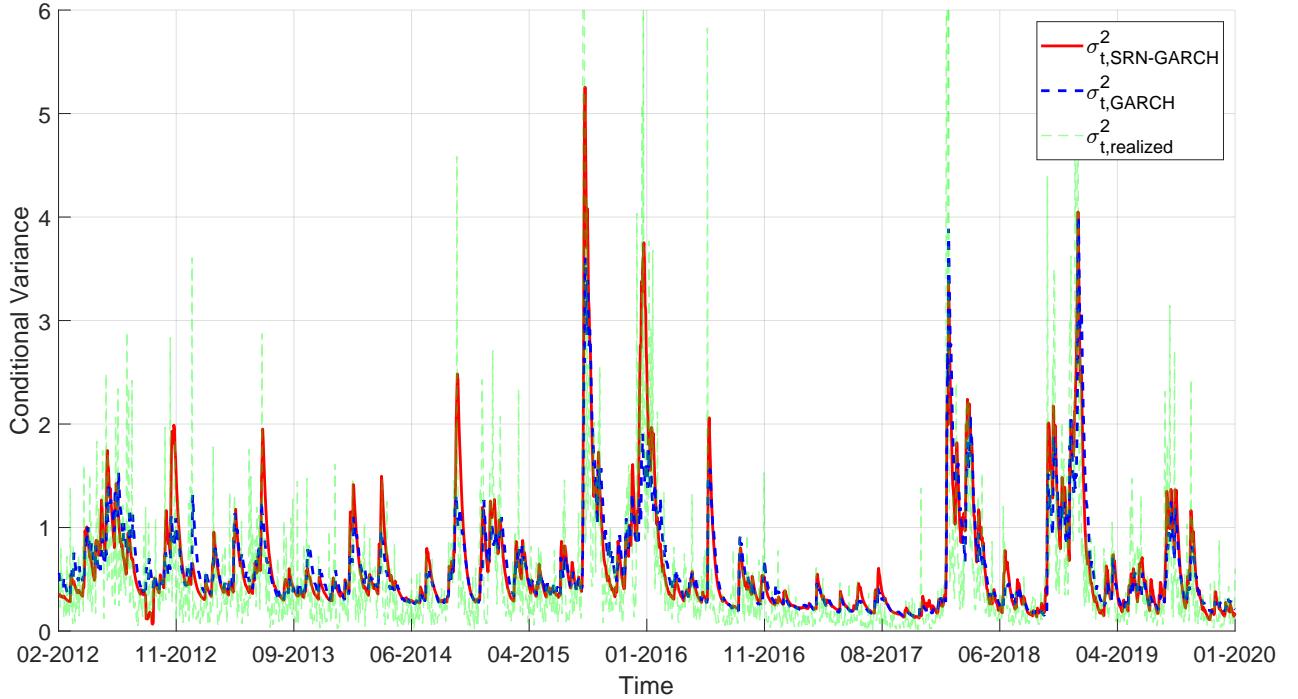


Figure 9: SP500 data: Forecast conditional variance by the GARCH (dashed) and SRN-GARCH (solid) models, together with the realized variance (dotted). (The figure is better viewed in colour).

are observed for these datasets. We note from Figures 9, 14 and 15 that in general, the RECH models and their GARCH-type counterparts produce forecast variances that adequately track the movement of the realized variance. The forecast variance of the RECH and benchmark models are similar in the low volatility regions while the RECH models have higher variance forecasts during high volatility periods. The variance forecasts of the RECH models seem to track the realized variance better than those of the benchmark GARCH-type models.

The in-sample analysis suggests that the RECH models fit the in-sample data of the four index datasets better than the counterpart GARCH-type models. However, it is possible that the superior in-sample performance is the result of overfitting (Pagan and Schwert, 1990; Donaldson and Kamstra, 1997). Table 16 provides summary statistics on the one-step-ahead forecasts of conditional variance and standardized residuals. The most important conclusion from Table 16 is that the RECH models do not overfit the data, as the forecast conditional variance of the RECH models are not excessively variable and the forecast residuals of the RECH models are very close to those of the GARCH-type benchmark models. The RECH models occasionally produce one-step-ahead forecast residuals with lower kurtosis than the counterpart GARCH-type models.

Table 17 shows the forecast performance of the models using the four predictive scores PPS, #Vio, QS and %Hit. As these four predictive scores complement each other, for each pair of the RECH and GARCH-type models, we compare their forecast performance by counting the number of times one model has a better predictive score than the other and report this count in the last column of Table 17. The model with the higher count is preferred. Table 17 shows

	Forecast Conditional Variance				Forecast Residual $\hat{\epsilon}_t$			
	Mean	Std	Skew	Kurtosis	Std	Skew	Kurtosis	LB- $\hat{\epsilon}_t$
SP500								
GARCH	0.597	0.476	2.805	14.391	0.932	-0.818	6.438	0.471
SRN-GARCH	0.620	0.546	2.309	9.009	0.921	-0.697	5.243	0.592
GJR	0.566	0.417	2.856	14.351	0.924	-0.738	5.775	0.484
SRN-GJR	0.626	0.565	2.443	9.868	0.932	-0.730	5.477	0.537
EGARCH	0.632	0.566	2.404	10.518	0.914	-0.815	5.939	0.438
SRN-EGARCH	0.634	0.615	2.945	13.554	0.910	-0.749	5.592	0.598
N225								
GARCH	0.930	1.061	4.553	30.093	0.977	-1.327	13.908	0.184
SRN-GARCH	0.884	0.921	3.752	21.485	1.004	-1.640	14.399	0.089
GJR	0.903	1.046	5.145	38.668	0.986	-1.495	15.828	0.128
SRN-GJR	0.889	0.944	3.761	21.200	1.006	-1.661	18.781	0.094
EGARCH	0.917	1.245	6.677	69.000	1.007	-1.599	17.672	0.146
SRN-EGARCH	0.897	1.231	6.858	74.418	1.017	-1.668	18.652	0.116
RUT								
GARCH	0.928	0.459	1.961	9.011	0.959	-0.471	4.101	0.721
SRN-GARCH	0.953	0.531	2.139	9.850	0.947	-0.547	4.378	0.562
GJR	0.919	0.448	2.034	8.473	0.953	-0.462	4.084	0.689
SRN-GJR	0.958	0.548	2.258	10.044	0.940	-0.465	4.080	0.572
EGARCH	0.990	0.545	1.828	8.450	0.932	-0.492	4.273	0.631
SRN-EGARCH	0.964	0.594	3.273	12.528	0.937	-0.462	4.106	0.618
DAX								
GARCH	0.867	0.474	1.238	4.394	0.973	-0.245	4.923	0.152
SRN-GARCH	0.904	0.585	1.566	5.807	0.962	-0.235	4.992	0.125
GJR	0.829	0.440	1.661	6.735	0.977	-0.235	4.733	0.097
SRN-GJR	0.907	0.594	1.646	6.226	0.964	-0.250	4.969	0.107
EGARCH	0.924	0.617	1.584	6.046	0.967	-0.296	5.205	0.124
SRN-EGARCH	0.902	0.634	1.987	7.982	0.972	-0.228	4.956	0.105

Table 16: Application: Summary statistics on the one-step-ahead out-of-sample forecast conditional variances $\hat{\sigma}_t^2$ and residual $\hat{\epsilon}_t$. The LB p-values denote the p-value from the Ljung-Box test with 10 lags.

that the RECH models consistently outperform their counterpart GARCH-type models for the S&P500, RUT and DAX data. For the N225 index, the predictive improvement of the RECH models over the benchmark counterparts is less clear, especially for the SRN-GARCH and GARCH models.

Tables 18 to 21 summarize the forecast performance measured by the predictive scores defined in Table 5. Each table contains six panels, corresponding to the six realized measures mentioned earlier. For each pair of the RECH and GARCH-type models, their forecast performance are also compared in the same way as in Table 17. Additionally, in each panel, bold numbers are used to indicate the lowest forecast errors. For each type of realized measure, the

	PPS	# Violation	QS	Hit Per.	Count
SP500					
GARCH	0.993	32	0.026	0.018	0
SRN-GARCH*	0.955	25	0.024	0.017	4
GJR	0.981	27	0.025	0.018	0
SRN-GJR*	0.959	25	0.024	0.017	4
EGARCH	0.963	29	0.025	0.018	0
SRN-EGARCH*	0.963	23	0.025	0.015	2
N225					
GARCH*	1.214	32	0.036	0.016	2
SRN-GARCH	1.216	33	0.035	0.016	1
GJR	1.217	31	0.036	0.017	1
SRN-GJR*	1.216	32	0.035	0.017	2
EGARCH	1.212	32	0.035	0.017	0
SRN-EGARCH*	1.212	30	0.035	0.017	1
RUT					
GARCH	1.298	34	0.032	0.018	0
SRN-GARCH*	1.286	26	0.030	0.016	4
GJR	1.290	31	0.031	0.017	0
SRN-GJR*	1.283	24	0.030	0.015	4
EGARCH	1.285	23	0.030	0.014	0
SRN-EGARCH*	1.281	23	0.030	0.014	1
DAX					
GARCH	1.257	47	0.031	0.020	0
SRN-GARCH*	1.247	38	0.029	0.020	3
GJR	1.249	50	0.030	0.022	0
SRN-GJR*	1.246	41	0.029	0.022	3
EGARCH	1.246	39	0.028	0.018	1
SRN-EGARCH*	1.243	37	0.029	0.020	3

Table 17: Applications: Forecast performance of the RECH and benchmark GARCH-type models. For the QS and %Hit measures, the results are calculated at the 1%-quantile. For each pair of the RECH and GARCH-type models, the asterisks indicate the model with the higher count.

model with the highest number of lowest forecast errors is preferred. The table shows that the RECH models in general outperform their counterpart GARCH-type models.

In particular, for the SP500, N225 and RUT datasets, the RECH models consistently perform the best across all panels. For example, the forecast results in Table 18 show that the SRN-GARCH model has the highest numbers of lowest forecast errors in all panels, implying that the SRN-GARCH model forecasts volatility the best for the SP500 data. For the N225 and RUT datasets, the SRN-EGARCH model clearly outperforms the other RECH and benchmark models in all realized measures. The superior predictive performance of the RECH models over the GARCH-type counterparts provides further evidence to support the conclusion that the RECH models do not overfit the index datasets.

Estimator		MSE ₁	MSE ₂	MAE ₁	MAE ₂	Qlike	R ² LOG	Count
BV	GARCH	0.088	0.780	0.220	0.351	0.115	0.715	0
	SRN-GARCH*	0.075	0.703	0.203	0.325	0.078	0.631	6
	GJR	0.077	0.721	0.206	0.322	0.097	0.678	1
	SRN-GJR*	0.076	0.691	0.205	0.328	0.080	0.642	5
RKV ₁	EGARCH	0.078	0.695	0.207	0.332	0.091	0.644	1
	SRN-EGARCH*	0.075	0.694	0.204	0.326	0.081	0.646	5
	GARCH	0.097	0.600	0.237	0.367	0.127	0.911	0
	SRN-GARCH*	0.084	0.538	0.222	0.346	0.093	0.815	6
RKV ₂	GJR	0.085	0.545	0.225	0.339	0.111	0.873	1
	SRN-GJR*	0.084	0.531	0.224	0.348	0.094	0.827	5
	EGARCH	0.086	0.528	0.224	0.349	0.098	0.810	2
	SRN-EGARCH	0.085	0.538	0.224	0.349	0.097	0.834	2
RKV ₃	GARCH	0.071	0.422	0.200	0.315	0.135	0.575	0
	SRN-GARCH*	0.061	0.371	0.187	0.296	0.109	0.507	6
	GJR	0.062	0.374	0.188	0.288	0.121	0.542	2
	SRN-GJR*	0.061	0.366	0.189	0.299	0.110	0.517	3
MedRV	EGARCH	0.063	0.368	0.190	0.301	0.117	0.514	2
	SRN-EGARCH*	0.062	0.376	0.188	0.298	0.111	0.521	4
	GARCH	0.070	0.401	0.199	0.314	0.137	0.569	0
	SRN-GARCH*	0.060	0.351	0.187	0.295	0.110	0.501	6
RV	GJR	0.060	0.353	0.187	0.287	0.123	0.536	2
	SRN-GJR*	0.060	0.347	0.188	0.297	0.112	0.511	3
	EGARCH	0.063	0.348	0.189	0.300	0.120	0.509	2
	SRN-EGARCH*	0.061	0.358	0.188	0.298	0.114	0.515	4
GARCH	GARCH	0.102	0.637	0.246	0.386	0.083	0.946	0
	SRN-GARCH*	0.087	0.555	0.227	0.357	0.039	0.849	6
	GJR	0.091	0.586	0.234	0.360	0.067	0.909	1
	SRN-GJR*	0.088	0.544	0.230	0.361	0.043	0.863	5
EGARCH	EGARCH	0.091	0.550	0.232	0.366	0.051	0.863	1
	SRN-EGARCH*	0.087	0.540	0.229	0.359	0.045	0.867	5
	GARCH	0.096	1.124	0.226	0.361	0.105	0.782	0
	SRN-GARCH*	0.083	1.054	0.212	0.340	0.073	0.701	6
SRN-GARCH	GJR	0.084	1.061	0.214	0.333	0.091	0.751	1
	SRN-GJR*	0.084	1.039	0.214	0.343	0.075	0.713	3
	EGARCH*	0.085	1.037	0.214	0.345	0.075	0.695	3
	SRN-EGARCH	0.084	1.045	0.214	0.344	0.078	0.720	2

Table 18: SP500 data: Forecast performance of the RECH and benchmark GARCH-type models using different realized measures. For each pair of the RECH and GARCH-type models, the asterisks indicate the model with the higher count. In each panel, the bold numbers indicate the best predictive scores.

As mentioned in Section 3.2.2, we now discuss an useful feature of the RECH models; that is, the volatility estimates and volatility forecasts of the RECH specifications are less sensitive to the choice of the structure for the GARCH component than a single GARCH-

Estimator		MSE ₁	MSE ₂	MAE ₁	MAE ₂	QLIKE	R ² LOG	Count
BV	GARCH	0.138	1.924	0.242	0.526	0.562	0.537	0
	SRN-GARCH*	0.124	1.795	0.227	0.487	0.557	0.491	6
	GJR	0.133	1.896	0.234	0.505	0.561	0.526	0
	SRN-GJR*	0.125	1.792	0.227	0.491	0.553	0.485	6
	EGARCH	0.130	1.870	0.226	0.499	0.553	0.478	0
	SRN-EGARCH*	0.124	1.794	0.224	0.487	0.552	0.472	6
RKV ₁	GARCH	0.216	3.792	0.311	0.661	0.573	1.028	0
	SRN-GARCH*	0.200	3.635	0.297	0.624	0.568	0.956	6
	GJR	0.210	3.758	0.304	0.641	0.574	1.012	0
	SRN-GJR*	0.201	3.648	0.298	0.628	0.565	0.952	6
	EGARCH	0.206	3.761	0.295	0.634	0.567	0.939	0
	SRN-EGARCH*	0.200	3.668	0.293	0.623	0.564	0.930	6
RKV ₂	GARCH	0.121	1.421	0.228	0.492	0.580	0.460	0
	SRN-GARCH*	0.108	1.268	0.215	0.455	0.575	0.420	6
	GJR	0.116	1.404	0.221	0.473	0.577	0.449	0
	SRN-GJR*	0.108	1.277	0.214	0.457	0.569	0.411	6
	EGARCH	0.114	1.438	0.213	0.466	0.569	0.405	0
	SRN-EGARCH*	0.108	1.304	0.211	0.454	0.569	0.400	5
RKV ₃	GARCH	0.120	1.405	0.228	0.491	0.581	0.456	0
	SRN-GARCH*	0.107	1.252	0.214	0.453	0.576	0.415	6
	GJR	0.115	1.387	0.220	0.471	0.578	0.444	0
	SRN-GJR*	0.107	1.261	0.213	0.455	0.569	0.407	6
	EGARCH	0.113	1.421	0.212	0.465	0.571	0.401	0
	SRN-EGARCH*	0.107	1.287	0.210	0.452	0.569	0.396	6
MedRV	GARCH	0.132	1.575	0.252	0.518	0.541	0.655	0
	SRN-GARCH*	0.118	1.415	0.238	0.484	0.530	0.600	6
	GJR	0.129	1.574	0.248	0.507	0.540	0.651	0
	SRN-GJR*	0.118	1.421	0.238	0.486	0.525	0.593	6
	EGARCH	0.125	1.636	0.238	0.502	0.524	0.580	0
	SRN-EGARCH*	0.118	1.458	0.234	0.482	0.523	0.576	6
RV	GARCH	0.145	2.068	0.246	0.535	0.577	0.557	0
	SRN-GARCH*	0.132	1.918	0.233	0.500	0.572	0.510	6
	GJR	0.141	2.046	0.239	0.515	0.576	0.546	0
	SRN-GJR*	0.133	1.927	0.232	0.501	0.567	0.503	6
	EGARCH	0.138	2.067	0.231	0.508	0.569	0.496	0
	SRN-EGARCH*	0.133	1.948	0.229	0.498	0.567	0.490	6

Table 19: N225 data: Forecast performance of the RECH and benchmark GARCH-type models using different realized measures. For each pair of the RECH and GARCH-type models, the asterisks indicate the model with the higher count. In each panel, the bold numbers indicate the best predictive scores.

type model. For example, for each dataset in Table 13 and 14, we compute the difference between the highest and lowest marginal likelihood estimates among the RECH specifications and calculate the same value for the GARCH-type benchmark models. Table 22 shows that

Estimator		MSE ₁	MSE ₂	MAE ₁	MAE ₂	QLIKE	R ² LOG	Count
BV	GARCH	0.099	0.593	0.246	0.468	0.708	0.534	0
	SRN-GARCH*	0.087	0.509	0.232	0.440	0.687	0.482	6
	GJR	0.089	0.538	0.236	0.443	0.694	0.499	0
	SRN-GJR*	0.087	0.509	0.233	0.442	0.686	0.686	6
	EGARCH	0.092	0.527	0.242	0.460	0.693	0.502	0
	SRN-EGARCH*	0.085	0.505	0.231	0.436	0.684	0.475	6
RKV ₁	GARCH	0.121	0.620	0.278	0.518	0.740	0.713	0
	SRN-GARCH*	0.108	0.543	0.266	0.493	0.720	0.651	6
	GJR	0.110	0.569	0.268	0.497	0.725	0.669	0
	SRN-GJR*	0.109	0.549	0.266	0.497	0.718	0.652	5
	EGARCH	0.113	0.562	0.273	0.510	0.725	0.670	0
	SRN-EGARCH*	0.106	0.540	0.263	0.491	0.714	0.640	6
RKV ₂	GARCH	0.097	0.524	0.247	0.464	0.708	0.565	0
	SRN-GARCH*	0.087	0.459	0.234	0.437	0.690	0.514	6
	GJR	0.090	0.483	0.238	0.441	0.697	0.537	0
	SRN-GJR*	0.088	0.464	0.235	0.440	0.690	0.520	6
	EGARCH	0.092	0.474	0.241	0.451	0.694	0.534	0
	SRN-EGARCH*	0.087	0.462	0.233	0.435	0.688	0.512	6
RKV ₃	GARCH	0.095	0.520	0.244	0.459	0.709	0.538	0
	SRN-GARCH*	0.085	0.456	0.231	0.433	0.691	0.489	6
	GJR	0.088	0.478	0.234	0.436	0.698	0.509	0
	SRN-GJR*	0.086	0.460	0.232	0.436	0.691	0.494	5
	EGARCH	0.090	0.472	0.238	0.448	0.696	0.508	0
	SRN-EGARCH*	0.084	0.459	0.230	0.431	0.690	0.487	6
MedRV	GARCH	0.137	0.950	0.289	0.553	0.669	0.819	0
	SRN-GARCH*	0.123	0.824	0.275	0.524	0.642	0.763	6
	GJR	0.127	0.879	0.278	0.527	0.654	0.784	0
	SRN-GJR*	0.123	0.817	0.276	0.525	0.641	0.764	6
	EGARCH	0.128	0.839	0.284	0.543	0.645	0.784	0
	SRN-EGARCH*	0.119	0.793	0.274	0.519	0.638	0.758	6
RV	GARCH	0.100	0.562	0.250	0.475	0.716	0.546	0
	SRN-GARCH*	0.089	0.490	0.235	0.443	0.698	0.492	6
	GJR	0.092	0.513	0.239	0.449	0.704	0.510	0
	SRN-GJR*	0.090	0.492	0.236	0.445	0.697	0.494	6
	EGARCH	0.094	0.507	0.244	0.462	0.704	0.513	0
	SRN-EGARCH*	0.088	0.492	0.234	0.441	0.695	0.486	6

Table 20: RUT data: Forecast performance of the RECH and benchmark GARCH-type models using different realized measures. For each pair of the RECH and GARCH-type models, the asterisks indicate the model with the higher count. In each panel, the bold numbers indicate the best predictive scores.

these discrepancies of in-sample performance among the RECH models are much smaller than those of the GARCH-type models, across all datasets. For each panel in Tables 18 to 21, we compute the difference between the highest and lowest forecast scores among the RECH

Estimator		MSE ₁	MSE ₂	MAE ₁	MAE ₂	QLIKE	R ² LOG	Count
BV	GARCH	0.083	0.572	0.216	0.404	0.618	0.449	0
	SRN-GARCH*	0.078	0.542	0.211	0.399	0.611	0.419	6
	GJR*	0.075	0.542	0.205	0.379	0.609	0.417	4
		0.078	0.545	0.210	0.399	0.606	0.412	2
	EGARCH	0.078	0.547	0.210	0.401	0.602	0.408	1
		0.077	0.548	0.207	0.395	0.601	0.401	5
	GARCH	0.103	0.691	0.245	0.452	0.628	0.609	0
	SRN-GARCH*	0.098	0.672	0.240	0.448	0.621	0.567	6
RKV ₁	GJR*	0.094	0.664	0.234	0.426	0.619	0.570	4
		0.098	0.676	0.239	0.448	0.616	0.560	2
	EGARCH	0.099	0.678	0.240	0.452	0.613	0.556	1
		0.097	0.681	0.237	0.447	0.612	0.549	5
RKV ₂	GARCH	0.067	0.396	0.197	0.368	0.623	0.354	0
	SRN-GARCH*	0.064	0.375	0.192	0.363	0.621	0.333	6
	GJR*	0.060	0.369	0.185	0.341	0.614	0.324	4
		0.064	0.379	0.191	0.363	0.615	0.326	2
	EGARCH	0.064	0.380	0.191	0.366	0.610	0.320	1
		0.063	0.382	0.188	0.359	0.610	0.314	4
RKV ₃	GARCH	0.068	0.410	0.198	0.370	0.623	0.362	0
	SRN-GARCH*	0.065	0.390	0.193	0.366	0.621	0.340	6
	GJR*	0.061	0.384	0.186	0.343	0.614	0.331	6
		0.065	0.394	0.193	0.367	0.615	0.333	0
	EGARCH	0.065	0.395	0.193	0.369	0.610	0.327	1
		0.064	0.397	0.189	0.363	0.610	0.321	4
MedRV	GARCH	0.084	0.525	0.223	0.416	0.586	0.490	0
	SRN-GARCH*	0.077	0.460	0.215	0.402	0.574	0.454	6
	GJR	0.077	0.490	0.213	0.392	0.578	0.462	2
		0.076	0.462	0.214	0.402	0.569	0.448	4
	EGARCH	0.076	0.458	0.213	0.403	0.562	0.436	1
		0.074	0.450	0.211	0.396	0.563	0.433	5
RV	GARCH	0.087	0.637	0.218	0.410	0.621	0.456	0
	SRN-GARCH*	0.083	0.619	0.213	0.407	0.616	0.426	6
	GJR*	0.079	0.613	0.206	0.384	0.613	0.423	4
		0.083	0.624	0.214	0.408	0.612	0.421	2
	EGARCH	0.083	0.624	0.214	0.412	0.608	0.415	1
		0.082	0.627	0.210	0.405	0.607	0.409	5

Table 21: DAX data: Forecast performance of the RECH and benchmark GARCH-type models using different realized measures. For each pair of the RECH and GARCH-type models, the asterisks indicate the model with the higher count. In each panel, the bold numbers indicate the best predictive scores.

models and do the same for the benchmark models; Figure 10 plots the results.

The comparison results in Figure 10 indicate that the discrepancies of out-of-sample performance among the RECH models are consistently lower than those of the GARCH-type

	SP500	N225	RUT	DAX
GARCH-type models	21.7	30.3	12.6	29.8
RECH models	6.9	6.8	5.5	3.3

Table 22: Applications: The difference between the highest and lowest marginal likelihood estimates of the RECH and the benchmark models across all in-sample data. The numbers are in the natural log scale.

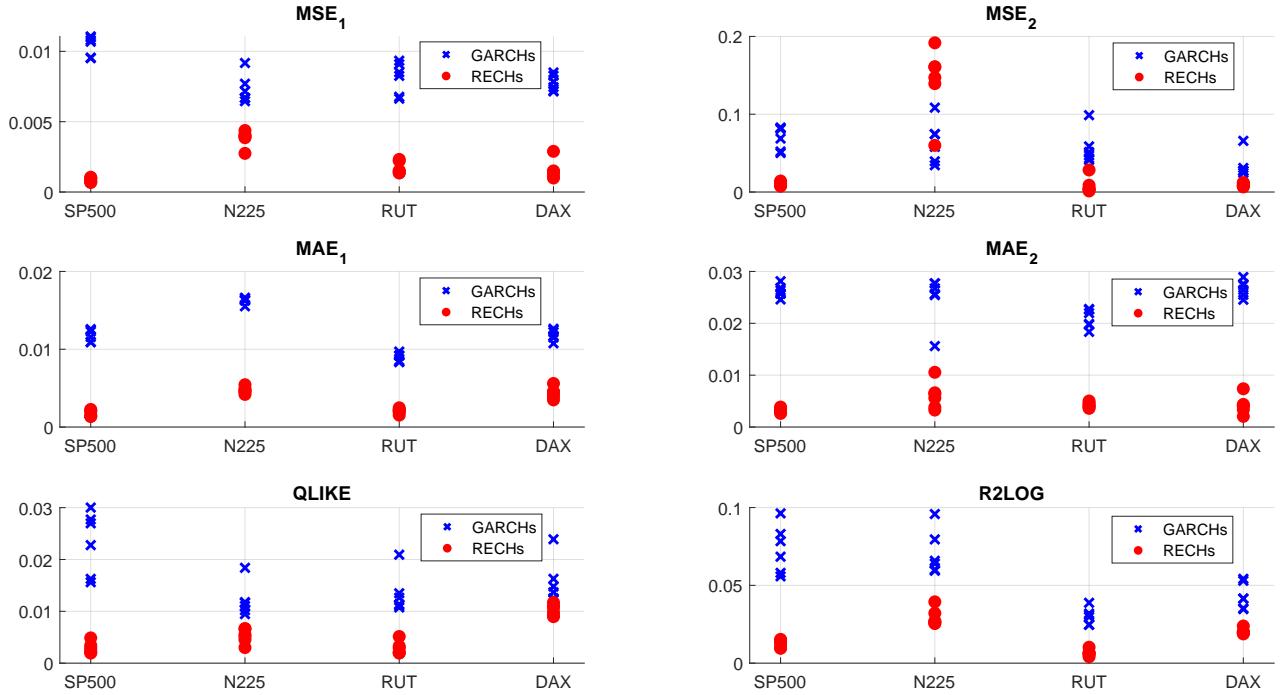


Figure 10: Applications: The difference between the highest and lowest forecast scores of the RECH and the benchmark models. In each panel, each column contains 6 values of the RECH models and 6 values of the benchmark models, corresponding to 6 realized measures.

models, except for the MSE_2 score for the N225 data. The result that model performance is less sensitive to the RECH specification is useful in practice as users do not need to worry about which specification should be used for their financial dataset.

5.3 Application to exchange rate data

This section reports on the application of the RECH model to analyse the USD/GBP daily exchange rates observed from 21/03/2001 to 01/03/2009⁴. We use the first 1000 observations for model estimation and the last 1000 observations for evaluating predictive performance.

The in-sample results (not shown) suggest that, unlike for the stock data, for exchange rate data GARCH performs the best compared to GJR and EGARCH. This is consistent with the study of Hansen and Lunde (2005) who show that there is no evidence that GARCH is outperformed by more sophisticated models on the DM/USD exchange rates. The in-sample

⁴The dataset was also downloaded from the Realized Library of The Oxford-Man Institute

result also shows that SRN-GARCH outperforms GARCH, thus being the best model, in terms of marginal likelihood.

	PPS	#Vio	QS	%Hit	MSE ₁	MSE ₂	MAE ₁	MAE ₂	QLike	R ² Log
GARCH	0.827	17	0.018	0.017	0.029	0.228	0.099	0.169	-0.121	0.166
SRN-GARCH*	0.827	14	0.017	0.017	0.024	0.180	0.094	0.154	-0.128	0.159
GJR	0.829	17	0.018	0.017	0.029	0.225	0.100	0.169	-0.121	0.167
SRN-GJR*	0.829	15	0.017	0.017	0.024	0.178	0.094	0.154	-0.127	0.160
EGARCH	0.836	17	0.019	0.016	0.039	0.297	0.112	0.196	-0.097	0.198
SRN-EGARCH*	0.843	15	0.017	0.017	0.028	0.194	0.107	0.178	-0.104	0.202

Table 23: USD/GBP exchange rate: one-step-ahead forecast comparison. The bold numbers denote the best scores. For each pair of the RECH and GARCH-type models, the asterisk indicates the models having better forecast performance.

Table 23 summarizes the forecast performance measured by the predictive scores discussed in Section 5, which suggests that RECH models are able to improve on their counterpart GARCH-type models in terms of volatility forecasts.

6 Conclusion

We propose a new class of conditional heteroskedastic models, which we call RECH models, by incorporating a RNN structure into the conditional variance of the GARCH-type models, and study in detail three RECH specifications: SRN-GARCH, SRN-GJR and SRN-EGARCH. We use Sequential Monte Carlo with likelihood annealing and data annealing for in-sample Bayesian inference and out-of-sample forecasting. We also use the estimate of marginal likelihood as a by-product of the SMC for model choice. The extensive simulation and empirical studies suggest that the RECH models not only have both attractive in-sample performance and accurate out-of-sample forecasts, but can also explain the volatility movement. In addition to the empirical study reported in Section 5, we tested the RECH models on all of the other datasets included in the Realized Library which contains 31 major stock markets around the world. In all cases, with the adding of the RNN component, the RECH model is not worse than its GARCH counterpart, whereas in some of these cases, significant predictive improvement is achieved by RECH.

An attractive feature of the proposed hybrid framework is that it is easy to use advances in both the deep learning and volatility modeling literatures to extend the current RECH models. This opens up many interesting future applications and areas of research. For example, one can use the Fourier Recurrent Unit (FRU) of Zhang et al. (2018) to construct the recurrent component of the RECH framework; FRU is currently considered as the state-of-the-art RNN architecture in deep learning. For the GARCH component, one can use the Bad Environment - Good Environment (BEGE) model of Bekaert et al. (2015), which can efficiently simulate the heavy tailed behavior of financial returns. Another interesting research direction is extending the univariate RECH models to the multivariate case. We conjecture that the recurrent neural network architectures will be more powerful for multivariate inputs as they can naturally capture the interaction between the inputs. This research is in progress.

Appendix

A1: SMC with data annealing

Algorithm 2 SMC with data annealing

1. Sample $\theta_0^j \sim p(\theta)$ and set $W_0^j = 1/M$ for $j = 1 \dots M$
2. For $t = 1, \dots, T$,

Step 1, reweighting: Compute the unnormalized weights

$$w_t^j = W_{t-1}^j p(y_t | y_{1:t-1}, \theta_{t-1}^j), \quad j = 1, \dots, M, \quad (23)$$

and set the new normalized weights as

$$W_t^j = \frac{w_t^j}{\sum_{s=1}^M w_s^s}, \quad j = 1, \dots, M. \quad (24)$$

Step 2: Compute the effective sample size (ESS)

$$\text{ESS} = \frac{1}{\sum_{j=1}^M (W_t^j)^2}. \quad (25)$$

if $\text{ESS} < cM$ for some $0 < c < 1$, **then**

- (i) **Resampling:** Resample from $\{\theta_{t-1}^j, W_t^j\}_{j=1}^M$, and then set $W_t^j = 1/M$ for $j = 1 \dots M$, to obtain the new equally-weighted particles $\{\theta_t^j, W_t^j\}_{j=1}^M$.
- (ii) **Markov move:** for each $j = 1, \dots, M$, move the sample θ_t^j according to N_{data} random walk Metropolis-Hastings steps:
 - (a) Generate a proposal $\theta_t^{j'}$ from multivariate normal distribution $\mathcal{N}(\theta_t^j, \Sigma_t)$ with Σ_t the covariance matrix.
 - (b) Set $\theta_t^j = \theta_t^{j'}$ with the probability

$$\min \left(1, \frac{p(y_{1:t} | \theta_t^{j'}) p(\theta_t^{j'})}{p(y_{1:t} | \theta_t^j) p(\theta_t^j)} \right) \quad (26)$$

otherwise keep θ_t^j .

end

A2: Additional results for Section 5

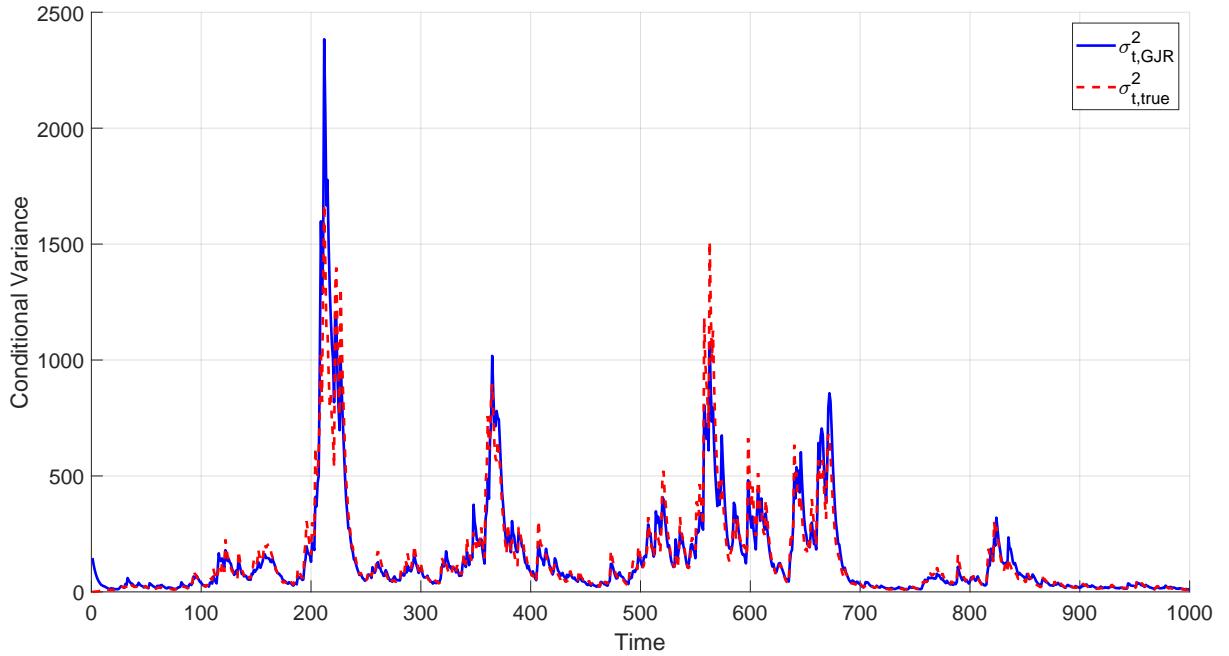


Figure 11: SIM II: The true conditional variance (dashed line) and estimated conditional variance (solid line) using the GJR model. (The figure is better viewed in colour).

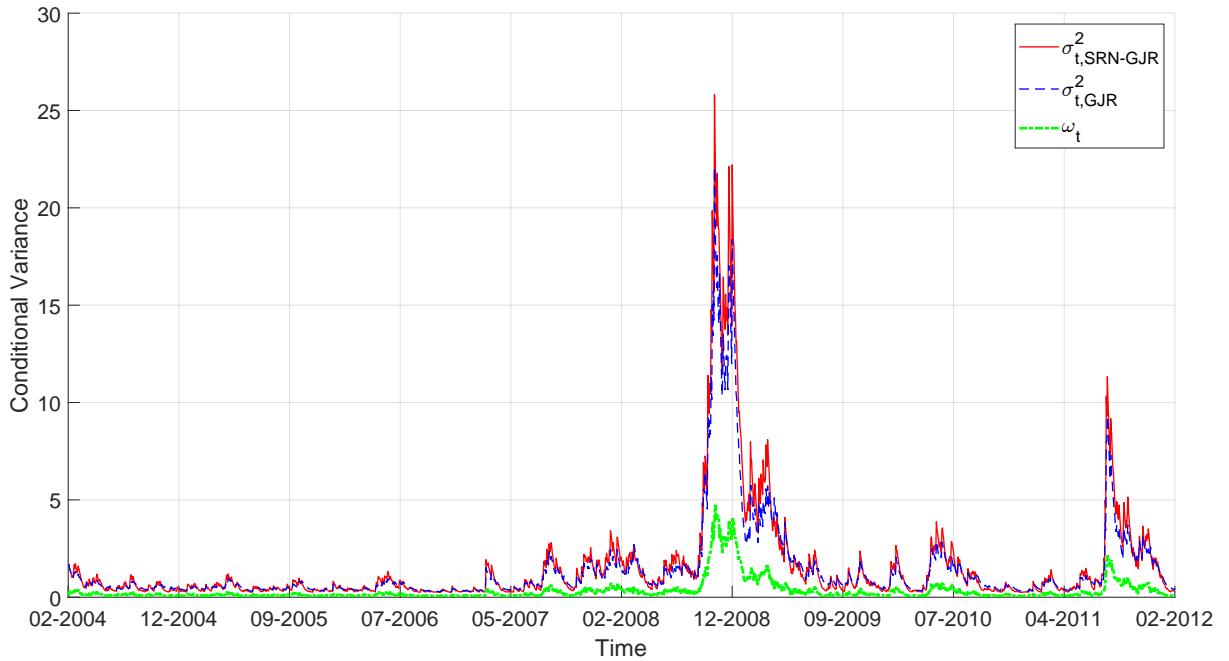


Figure 12: SP500: The in-sample conditional variance of the GJR (dashed line) and SRN-GJR (solid line) at all time steps. The bottom line shows the values of the recurrent component ω_t of the SRN-GJR specification. (The figure is better viewed in colour).

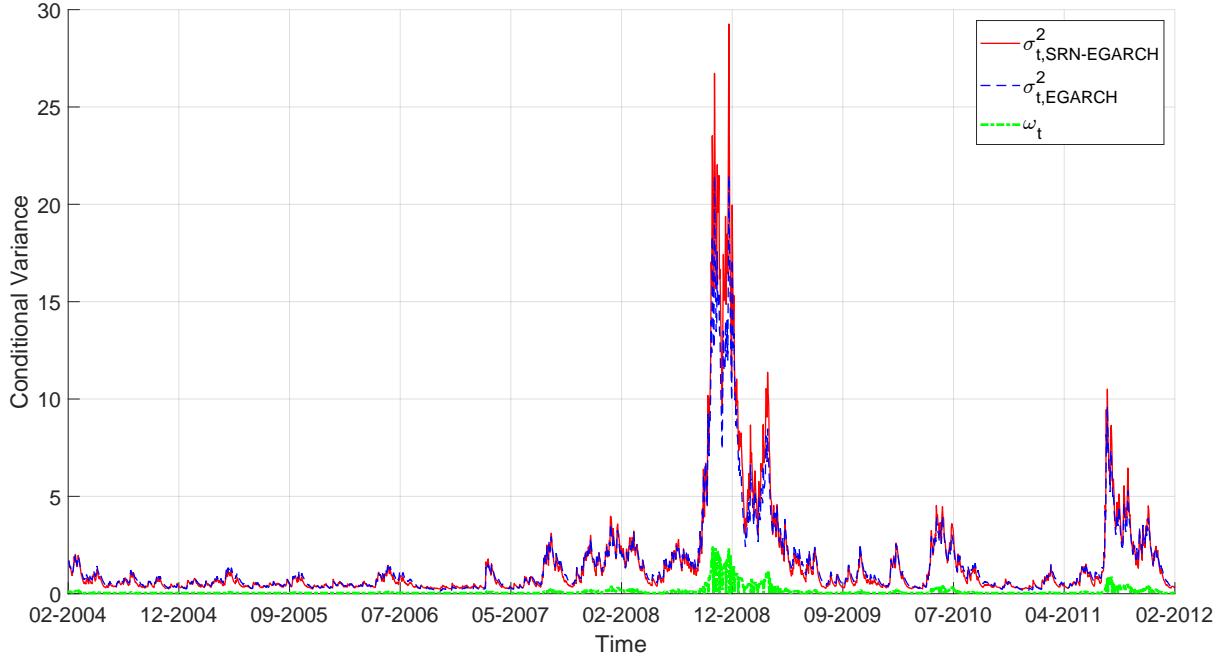


Figure 13: SP500: The in-sample conditional variance of the EGARCH (dashed line) and SRN-EGARCH (solid line) at all time steps. The bottom line shows the values of the recurrent component ω_t of the SRN-EGARCH specification. (The figure is better viewed in colour).

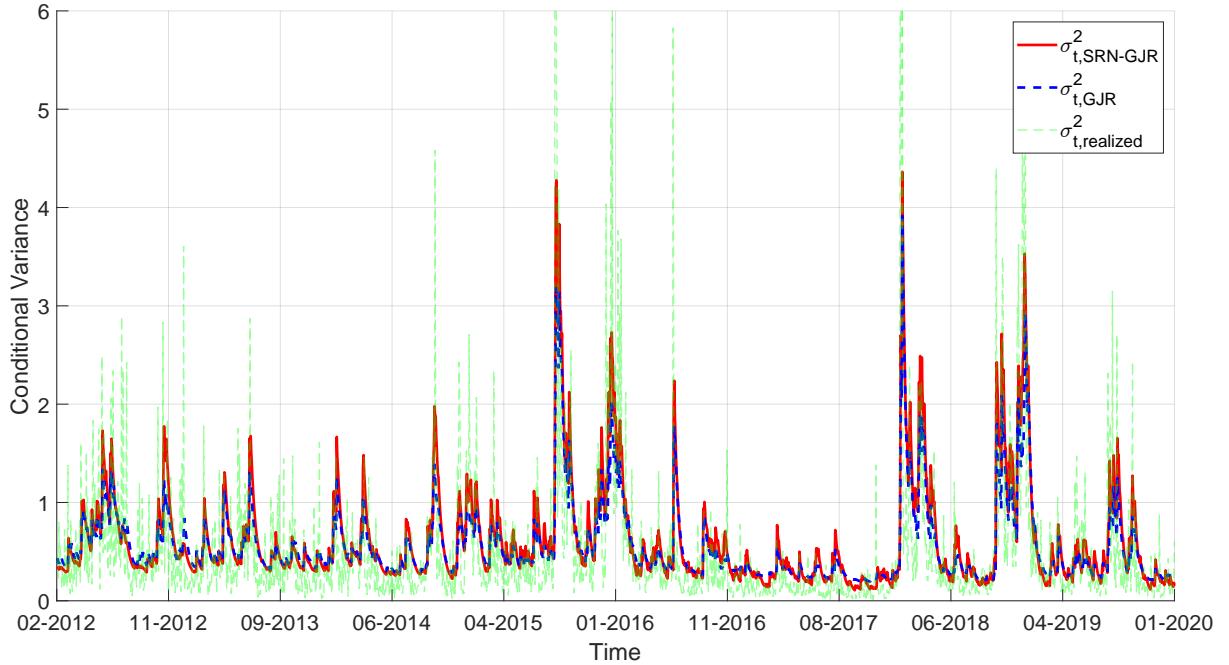


Figure 14: SP500: Forecast conditional variance by the GJR (dashed) and SRN-GJR (solid) models, together with the realized variance (dotted). (The figure is better viewed in colour).

A3: Comparison to the LSTM-SV and GP-Vol models

This section reports the comparison of the predictive performance between the RECH and other non-linear volatility models including the LSTM-SV model of Nguyen et al. (2019) and

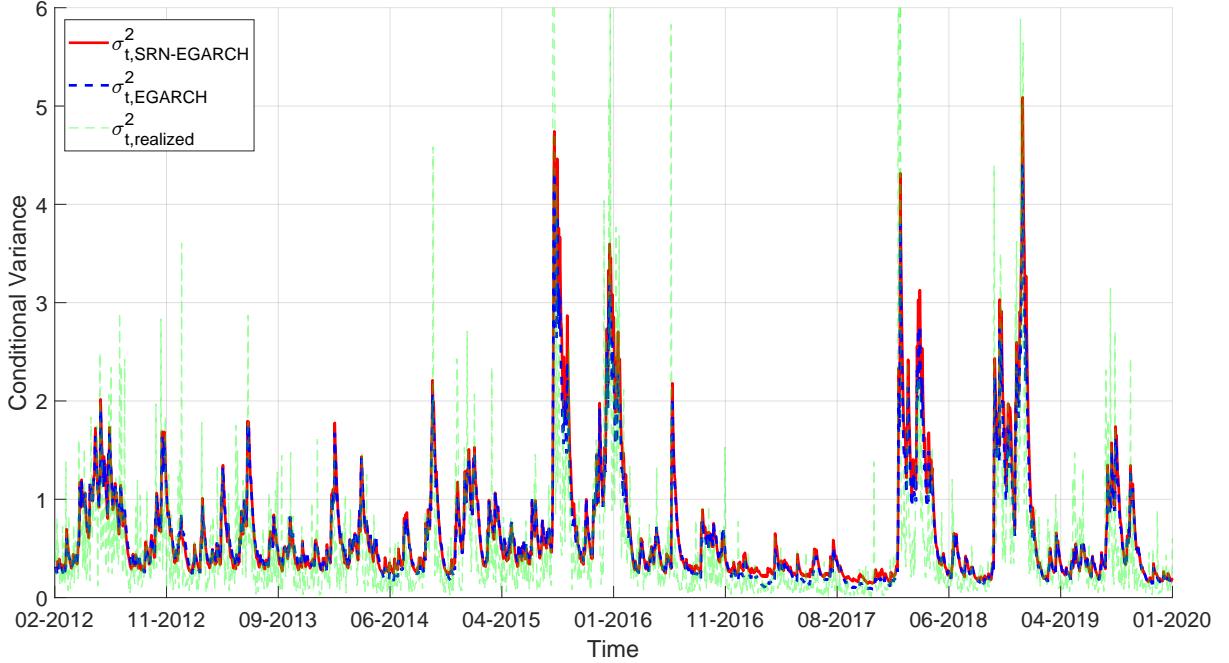


Figure 15: SP500: Forecast conditional variance by the EGARCH (dashed) and SRN-EGARCH (solid) models, together with the realized variance (dotted). (The figure is better viewed in colour).

the GP-Vol model of Wu et al. (2014). The GP-Vol model is an engineering-oriented stochastic volatility model using a Gaussian process to capture the non-linearity in the volatility dynamics. As the GP-Vol model focuses mainly on prediction, we use it as another benchmark model, together with the LSTM-SV model, to assess the predictive performance of the RECH models. We use the software packages provided by Nguyen et al. (2019) and Wu et al. (2014) to perform Bayesian inference and prediction for the LSTM-SV and GP-Vol models, respectively, with all settings at their default values. For all models, we use the posterior means estimated from in-sample data to perform one-step-ahead forecasting for out-of-sample data. We use $T_{\text{out}} = 1000$ observations for the out-of-sample period.

Table 24 shows the out-of-sample performance of the models evaluated on the SP500 index, using the same realize measures discussed in Section 5.2. Table 24 suggests that RECH models in general predict better than the LSTM-SV and GP-Vol models. The superiority of the SRN-GARCH model, and RECH models in general, over the LSTM-SV model is expected as the SRN component in RECH models is able to capture the leverage effect, while this is not the case for the LSTM component in LSTM-SV. We note that the GP-Vol model uses a Gaussian process with its covariance matrix expanding over time and hence it becomes computationally expensive in applications with long time series. We observe similar results between the RECH, LSTM-SV and GP-Vol models for the other datasets in Section 5.2.

Estimator		MSE ₁	MSE ₂	MAE ₁	MAE ₂	QLIKE	R ² LOG
BV	GP-Vol	0.176	1.586	0.309	0.555	0.546	1.097
	LSTM-SV	0.104	1.326	0.237	0.397	0.354	0.717
	SRN-GARCH*	0.105	1.210	0.201	0.332	0.448	0.572
	SRN-GJR*	0.096	1.210	0.220	0.370	0.288	0.637
RKV ₁	SRN-EGARCH	0.100	1.222	0.228	0.383	0.297	0.681
	GP-Vol	0.173	1.039	0.317	0.549	0.501	1.260
	LSTM-SV	0.115	0.829	0.263	0.425	0.366	0.978
	SRN-GARCH*	0.129	1.059	0.240	0.395	0.598	0.772
	SRN-GJR	0.104	0.875	0.240	0.400	0.369	0.785
RKV ₂	SRN-EGARCH	0.107	0.882	0.246	0.411	0.373	0.826
	GP-Vol	0.154	0.851	0.295	0.522	0.603	0.955
	LSTM-SV	0.078	0.524	0.213	0.349	0.383	0.567
	SRN-GARCH*	0.070	0.440	0.179	0.271	0.307	0.471
	SRN-GJR	0.077	0.440	0.209	0.332	0.224	0.587
RKV ₃	SRN-EGARCH	0.083	0.457	0.220	0.350	0.236	0.635
	GP-Vol	0.155	0.893	0.296	0.523	0.595	0.963
	LSTM-SV	0.080	0.574	0.214	0.352	0.379	0.575
	SRN-GARCH*	0.074	0.499	0.182	0.278	0.326	0.479
	SRN-GJR	0.077	0.480	0.208	0.332	0.232	0.579
MedRV	SRN-EGARCH	0.082	0.496	0.219	0.350	0.244	0.626
	GP-Vol	0.194	1.147	0.333	0.585	0.559	1.329
	LSTM-SV	0.117	0.830	0.266	0.437	0.336	0.936
	SRN-GARCH*	0.103	0.734	0.209	0.327	0.303	0.703
	SRN-GJR	0.106	0.639	0.253	0.400	0.196	0.927
RV	SRN-EGARCH	0.112	0.645	0.263	0.414	0.210	0.990
	GP-Vol	0.186	2.048	0.318	0.571	0.488	1.189
	LSTM-SV	0.123	1.865	0.255	0.429	0.342	0.861
	SRN-GARCH	0.143	2.366	0.228	0.404	0.337	0.675
	SRN-GJR*	0.114	2.132	0.227	0.406	0.381	0.658
	SRN-EGARCH	0.117	2.144	0.236	0.420	0.387	0.696

Table 24: SP500: Out-of-sample performance of the GP-Vol, LSTM-SV and RECH models using different realized measures. In each panel, the bold numbers indicate the best predictive scores and the asterisk indicates the model with the best predictive performance.

A4: Proofs

Proof of Theorem 1. Recall the σ -fields $\mathcal{F}_t = \sigma(y_s, s \leq t)$, $t \geq 1$, and let us define \mathcal{F}_0 to be the σ -field generated by σ_0^2 . As the recurrent component is bounded,

$$\mathbb{E}(y_t^2 | \mathcal{F}_{t-1}) = \sigma_t^2 \leq M + \alpha\sigma_{t-1}^2 + \beta y_{t-1}^2, \quad t > 1. \quad (27)$$

We have that

$$\begin{aligned}\mathbb{E}(y_t^2|\mathcal{F}_{t-2}) &= \mathbb{E}(\mathbb{E}(y_t^2|\mathcal{F}_{t-1})|\mathcal{F}_{t-2}) \\ &\leq M + \alpha\sigma_{t-1}^2 + \beta\mathbb{E}(y_{t-1}^2|\mathcal{F}_{t-2}) \\ &= M + \alpha\sigma_{t-1}^2 + \beta\sigma_{t-1}^2,\end{aligned}$$

hence, by (27),

$$\mathbb{E}(y_t^2|\mathcal{F}_{t-2}) \leq \begin{cases} M + (\alpha + \beta)\sigma_1^2 < M + \sigma_0^2, & t = 2 \\ M + (\alpha + \beta)(M + \alpha\sigma_{t-2}^2 + \beta y_{t-2}^2), & t > 2. \end{cases}$$

Similarly,

$$\begin{aligned}\mathbb{E}(y_t^2|\mathcal{F}_{t-3}) &= \mathbb{E}(\mathbb{E}(y_t^2|\mathcal{F}_{t-2})|\mathcal{F}_{t-3}) \\ &\leq M + (\alpha + \beta)(M + \alpha\sigma_{t-2}^2 + \beta\mathbb{E}(y_{t-2}^2|\mathcal{F}_{t-3})) \\ &= M(1 + (\alpha + \beta)) + (\alpha + \beta)^2\sigma_{t-2}^2, \quad t \geq 3.\end{aligned}$$

For $t=3$,

$$\mathbb{E}(y_t^2|\mathcal{F}_{t-3}) = M(1 + (\alpha + \beta)) + (\alpha + \beta)^2\sigma_1^2 < M(1 + (\alpha + \beta)) + \sigma_0^2,$$

and for $t>3$, by (27),

$$\begin{aligned}\mathbb{E}(y_t^2|\mathcal{F}_{t-3}) &\leq M(1 + (\alpha + \beta)) + (\alpha + \beta)^2(M + \alpha\sigma_{t-3}^2 + \beta y_{t-3}^2) \\ &\leq M(1 + (\alpha + \beta) + (\alpha + \beta)^2) + (\alpha + \beta)^2(\alpha\sigma_{t-3}^2 + \beta y_{t-3}^2).\end{aligned}$$

Hence

$$\mathbb{E}(y_t^2|\mathcal{F}_{t-3}) \leq \begin{cases} M(1 + (\alpha + \beta)) + \sigma_0^2, & t = 3 \\ M(1 + (\alpha + \beta) + (\alpha + \beta)^2) + (\alpha + \beta)^2(\alpha\sigma_{t-3}^2 + \beta y_{t-3}^2), & t > 3. \end{cases}$$

By deduction we have that, for $k=2,\dots,t-1$,

$$\mathbb{E}(y_t^2|\mathcal{F}_{t-k}) \leq \begin{cases} M(1 + \sum_{i=1}^{k-2}(\alpha + \beta)^i) + \sigma_0^2, & t = k \\ M(1 + \sum_{i=1}^{k-1}(\alpha + \beta)^i) + (\alpha + \beta)^{k-1}(\alpha\sigma_{t-k}^2 + \beta y_{t-k}^2), & t > k. \end{cases} \quad (28)$$

Therefore,

$$\begin{aligned}\mathbb{V}(y_t^2|\sigma_0^2) \leq \mathbb{E}(y_t^2|\mathcal{F}_0) &\leq M\left(1 + \sum_{i=1}^{k-2}(\alpha + \beta)^i\right) + \sigma_0^2 \\ &< M\left(1 + \sum_{i=1}^{\infty}(\alpha + \beta)^i\right) + \sigma_0^2 = \frac{M}{1 - \alpha - \beta} + \sigma_0^2.\end{aligned}$$

□

References

- Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4):885–905.
- Andersen, T. G., Dobrev, D., and Schaumburg, E. (2012). Jump-robust volatility estimation using nearest neighbor truncation. *Journal of Econometrics*, 169(1):75 – 93. Recent Advances in Panel Data, Nonlinear and Nonparametric Models: A Festschrift in Honor of Peter C.B. Phillips.
- Ardia, D. and Hoogerheide, L. F. (2010). Bayesian estimation of the GARCH(1,1) model with student- t innovations. *The R Journal*, 2(2):41–47.
- Baillie, R. T., Bollerslev, T., and Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 74(1):3 – 30.
- Barndorff-Nielsen, O., Hansen, P., Lunde, A., and Shephard, N. (2008). Designing realized kernels to measure the ex-post variation of equity prices in the presence of noise. *Econometrica*, 76(6):1481–1536.
- Barndorff-Nielsen, O. E. and Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2(1):1–37.
- Bekaert, G., Engstrom, E., and Ermolov, A. (2015). Bad environments, good environments: A non-Gaussian asymmetric volatility model. *Journal of Econometrics*, 186(1):258 – 275.
- Black, F. (1976). Studies of stock price volatility changes. In *Proceedings of the 1976 Meeting of the Business and Economic Statistics Section, American Statistical Association*, pages 177–181.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307 – 327.
- Bollerslev, T. (2008). Glossary to ARCH (GARCH). Creates research papers, Department of Economics and Business Economics, Aarhus University.
- Box, G. E. P. and Jenkins, G. (1976). *Time Series Analysis, Forecasting and Control*. Holden-Day, Inc., San Francisco, CA, USA.
- Breidt, F., Crato, N., and de Lima, P. (1998). The detection and estimation of long memory in stochastic volatility. *Journal of Econometrics*, 83(1):325 – 348.
- Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, 89(3):539–551.
- Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society, Series B*, 68:411–436.
- Donaldson, R. G. and Kamstra, M. (1997). An artificial neural network-GARCH model for international stock return volatility. *Journal of Empirical Finance*, 4(1):17–46.

- Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14:179–21.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4):987–1007.
- Fleming, J., Kirby, C., and Ostdiek, B. (2003). The economic value of volatility timing using “realized” volatility. *Journal of Financial Economics*, 67(3):473 – 509.
- Giraitis, L., Kokoszka, P., Leipus, R., and Teyssière, G. (2003). Rescaled variance and related tests for long memory in volatility and levels. *Journal of Econometrics*, 112(2):265 – 294.
- Glosten, L. R., Jagannathan, R., and 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*, 48(5):1779–1801.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.
- Greaves-Tunnell, A. and Harchaoui, Z. (2019). A statistical investigation of long memory in language and music. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2394–2403. PMLR.
- Gunawan, D., Dang, K., Quiroz, M., Kohn, R., and Tran, M. (2020). Subsampling sequential Monte Carlo for static Bayesian models. *Statistics and Computing*, 30:1741–1758.
- Hansen, P. R. and Lunde, A. (2005). A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, 20(7):873–889.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.
- Hosszejni, D. and Kastner, G. (2021). Modeling univariate and multivariate stochastic volatility in R with stochvol and factorstochvol. *Journal of Statistical Software (forthcoming)*.
- Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(2):203–222.
- Jeffreys, H. (1961). *Theory of Probability*, 3rd. Clarendon Press, Oxford, England.
- Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). Markov Chain Monte Carlo in practice: A roundtable discussion. *The American Statistician*, 52(2):93–100.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.
- Kim, H. Y. and Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications*, 103:25 – 37.

- Koopman, S. J., Lucas, A., and Scharth, M. (2016). Predicting time-varying parameters with parameter-driven and observation-driven models. *The Review of Economics and Statistics*, 98(1):97–110.
- Le, Q. V., Jaitly, N., and Hinton, G. E. (2015). A simple way to initialize recurrent networks of rectified linear units. *CoRR*, abs/1504.00941.
- Li, D., Clements, A., and Drovandi, C. (2020). Efficient Bayesian estimation for GARCH-type models via Sequential Monte Carlo. *Econometrics and Statistics*, 19.
- Liew, S. S., Khalil-Hani, M., and Bakhteri, R. (2016). Bounded activation functions for enhanced training stability of deep neural networks on visual pattern recognition problems. *Neurocomputing*, 216:718–734.
- Lipton, Z., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. arXiv:1804.04359.
- Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93(443):1032–1044.
- Liu, Y. (2019). Novel volatility forecasting using deep learning—long short term memory recurrent neural networks. *Expert Systems with Applications*, 132:99–109.
- Lo, A. W. (1991). Long-term memory in stock market prices. *Econometrica*, 59(5):1279–1313.
- Luo, R., Zhang, W., Xu, X., and Wan, J. (2018). A neural stochastic volatility model. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.
- Mandelbrot, B. (1967). The variation of some other speculative prices. *The Journal of Business*, 40(4):393–413.
- Martens, M. (2002). Measuring and forecasting S&P 500 index-futures volatility using high-frequency data. *Journal of Futures Markets*, 22(6):497–518.
- Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML’10, pages 807–814, USA. Omnipress.
- Neal, R. (2001). Annealed importance sampling. *Statistics and Computing*, 11:125–139.
- Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2):347–370.
- Nguyen, N., Tran, M.-N., Gunawan, D., and Kohn, R. (2019). A long short-term memory stochastic volatility model. *arXiv e-prints*, page arXiv:1906.02884.
- Oliva, J. B., Póczos, B., and Schneider, J. G. (2017). The statistical recurrent unit. In *ICML2017*.
- Pagan, A. R. and Schwert, G. (1990). Alternative models for conditional stock volatility. *Journal of Econometrics*, 45(1):267 – 290.

- Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11.
- Poon, S.-H. and Granger, C. W. (2003). Forecasting volatility in financial markets: A review. *Journal of Economic Literature*, 41(2):478–539.
- Roh, T. H. (2007). Forecasting the volatility of stock price index. *Expert Systems with Applications*, 33(4):916 – 922.
- Shephard, N. and Sheppard, K. (2010). Realising the future: forecasting with high-frequency-based volatility (heavy) models. *Journal of Applied Econometrics*, 25(2):197–231.
- Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. *Journal of Business & Economic Statistics*, 37(1):121–133.
- Taylor, S. (1986). *Modelling Financial Time Series*. John Wiley, Chichester.
- van Bellegem, S. (2012). Locally stationary volatility modeling. In Bauwens, L., Hafner, C., and Laurent, S., editors, *Volatility Models and Their Applications*. Wiley & Sons.
- Wu, Y., Hernández-Lobato, J. M., and Ghahramani, Z. (2014). Gaussian process volatility model. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.
- Zhang, G. (2003). Time series forecasting using a hybrid arima and neural network model. *Neurocomputing*, 50:159 – 175.
- Zhang, G., Patuwo, B. E., and Hu, M. Y. (1998). Forecasting with artificial neural networks:: The state of the art. *International Journal of Forecasting*, 14(1):35 – 62.
- Zhang, J., Lin, Y., Song, Z., and Dhillon, I. S. (2018). Learning long term dependencies via Fourier recurrent units. In *ICML2017*.