Combining Long Memory and Level Shifts in Modeling and Forecasting the Volatility of Asset Returns

Rasmus T. Varneskov† and Pierre Perron‡

Aarhus University and CREATES Boston University

This version: September 8, 2015

Abstract

We propose a parametric state space model of asset return volatility with an accompanying estimation and forecasting framework that allows for ARFIMA dynamics, random level shifts and measurement errors. The Kalman filter is used to construct the state-augmented likelihood function and subsequently to generate forecasts, which are mean- and path-corrected. We apply our model to eight daily volatility series constructed from both high-frequency and daily returns. Full sample parameter estimates reveal that random level shifts are present in all series. Genuine long memory is present in high-frequency measures of volatility whereas there is little remaining dynamics in the volatility measures constructed using daily returns. From extensive forecast evaluations, we find that our ARFIMA model with random level shifts consistently belongs to the 10% Model Confidence Set across a variety of forecast horizons, asset classes, and volatility measures. The gains in forecast accuracy can be very pronounced, especially at longer horizons.

Keywords: Forecasting, Kalman Filter, Long Memory Processes, State Space Modeling, Stochastic Volatility, Structural Change.

JEL classification: C13, C22, C53

---

*We wish to thank Asger Lunde for providing cleaned high frequency tick data. Furthermore, we thank Torben G. Andersen, Kim Christensen, Oliver Linton, Morten Ørregaard Nielsen, Anders Rahbek as well as seminar participants at CREATES, Aarhus University, and the 2nd Humboldt-Copenhagen conference, May 2011, for helpful comments. An earlier version of this paper entitled "Combining Long Memory and Level Shifts in Modeling and Forecasting of Persistent Time Series" has been circulated since November 2010. Financial support from Aarhus School of Business and Social Sciences, Aarhus University, and CREATES, funded by the Danish National Research Foundation, is gratefully acknowledged.

†Department of Economics and Business, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus V., Denmark. Email: rvarneskov@creates.au.dk.
‡Department of Economics, Boston University, 270 Bay State Rd., Boston, MA, 02215. Email: perron@bu.edu.
1 Introduction

The literature on asset return volatility modeling has surged since the introduction of the ARCH model by Engle (1982) due to numerous potential applications in financial economics such as asset- and derivative pricing, risk management and portfolio selection. In addition, various volatility-linked derivatives are nowadays being actively traded on the Chicago Board of Options Exchange and in over-the-counter markets. Recently, Andersen, Bollerslev, Diebold & Ebens (2001), Andersen, Bollerslev, Diebold & Labys (2001, 2003), Koopman, Jungbacker & Hol (2005), Deo, Hurvich & Lu (2006), Andersen, Bollerslev & Diebold (2007), Corsi (2009), Chiriac & Voev (2011) and Varneskov & Voev (2013), among others, demonstrate that various realized volatility time series display characteristics compatible with fractionally integrated, or $I(d)$, processes, and that the modeling of such “long memory” properties significantly improves the precision of out-of-sample forecasts of future return volatility.

We may formally define fractional integration or, as we will label it throughout, genuine long memory as follows; let $e_t = C(L)\epsilon_t$ with $\epsilon_t \sim i.i.d.(0, \sigma^2_\epsilon)$ and $\mathbb{E}(|\epsilon_t|^r) < \infty$ for some $r > 2$ be a short memory process with lag polynomial $C(L) = \sum_{i=0}^{\infty} c_i L^i$ satisfying $\sum_{i=0}^{\infty} i|c_i| < \infty$ and $C(1) \neq 0$, then $h_t = (1 - L)^d e_t$ for $t = 1, 2, \ldots$ is fractionally integrated of order $I(d)$ with autocovariance function

$$R_h(\tau) = g(\tau)\tau^{2d-1} \quad \text{as} \quad \tau \to \infty$$

where $g(\tau)$ is a slowly varying function as $\tau \to \infty$. The properties of such processes depend crucially on the magnitude of the fractional integration order $d$. In this paper, we shall mainly be concerned with the case $0 \leq d < 1/2$, i.e., with a stationary process that exhibits genuine long memory whenever $d > 0$, and which is characterized by having hyperbolically decaying autocovariances. However, we will also make references to the non-stationary case $d \geq 1/2$. The fractional ARIMA, or ARFIMA, model, independently introduced by Granger & Joyeux (1980) and Hosking (1981), is a flexible time series specification that captures genuine long memory and, as a result, has become popular for volatility modeling and forecasting, e.g., Andersen, Bollerslev, Diebold & Labys (2003).

Recently, however, a parallel literature has studied the possibility of genuine long memory being confused with a short memory process contaminated by random level shifts, spurred by the expositions in Perron (1989, 1990), who show that unit roots ($d = 1$) and structural changes are easily confused in the sense that the sum of the autoregressive coefficients is biased towards one if a stationary process is contaminated by level shifts. Applying this concept to the context of genuine long memory modeling, Lobato & Savin (1998), Diebold & Inoue (2001), Granger & Hyung (2004), and Perron & Qu (2007, 2010), among others, show theoretically and through simulations that if a short memory process is contaminated by random level shifts, the resulting time series will display many of the same characteristics as one of genuine long memory; for example, the hyperbolically decaying autocovariances.\footnote{Related findings are made by Bhattacharya, Gupta & Waymire (1983), Mikosch & Stărică (2004), Stărică & Granger (2005), and Ohanissian, Russell & Tsay (2008).}

Motivated by these findings, Lu & Perron (2010) and Qu & Perron (2013), extending earlier work by Chen & Tiao
(1990) and McCulloch & Tsay (1993), propose parametric models of asset return volatility, which allow for both random level shifts and short memory dynamics. They perform empirical analyses of daily stock index returns and argue that the (genuine) long memory properties of the series are, indeed, spurious. These findings are corroborated in Xu & Perron (2014). Similar conclusions arise from another branch of the literature, which consider semi-parametric estimation and testing for genuine long memory, see, e.g., Smith (2005), Perron & Qu (2010), Qu (2011), McCloskey & Perron (2013), and McCloskey & Hill (2014). However, the proposed semi-parametric frameworks have a disadvantage that random level shifts are not identified, making them unsuitable for forecasting.

As such, we face a dual problem. The presence of random level shifts may bias the parameter estimates for genuine long memory models, and, consequently, lead to misspecified dynamics of asset return volatility. However, the presence of genuine long memory may also cause spurious detection of random level shifts in the series, see, e.g., Nunes, Newbold & Kuan (1995) and Granger & Hyung (2004). As a solution to this problem, we propose a parametric framework for asset return volatility modeling, which allows volatility to exhibit both random level shifts and ARFIMA dynamics. Furthermore, we allow for measurement errors in the observable volatility proxies such that we may analyze series constructed from daily as well as high-frequency data. The idea of combining random level shifts with a fractionally integrated component for time series modeling resembles the strategy in Ray & Tsay (2002). However, we introduce a framework that augments their Bayesian approach in four different directions; by allowing for a short memory ARMA component, by allowing for measurement errors, by allowing random level shifts to occur at each time $t$, and not in blocks, and, finally, we extend their analysis by providing a novel forecasting framework.

In particular, we propose a parametric state space framework to estimate the model and perform out-of-sample forecasting. The estimation procedure is similar to the one introduced in Perron & Wada (2009) and Lu & Perron (2010) where the basic principle is to augment the probability of states by the realizations of a mixture of normally distributed processes and apply the Kalman filter to construct the likelihood function conditional on the realization of states. However, an additional challenge arises since there exists no exact finite state space representation if the underlying process contains a genuine long memory component. This problem is solved through truncation. Furthermore, to show the adequacy of the estimation methodology and to analyze the role of truncation order, we set up a simulation study where we compare the estimated memory parameters from our random level shift ARFIMA, or RLS-ARFIMA, model with standard ARFIMA parameter estimates, and we illustrate how the latter are affected by random level shifts. The recursive structure of the Kalman filter allows us to introduce a new forecasting framework for general parametric random level shift models, which utilizes the information in the Kalman recursions to generate forecasts for a given state and, then, weight them with the probability of being on a given transition path. Hence, the forecasts are both mean- and path-corrected.

We apply the proposed reduced form modeling framework to eight daily asset return volatility series, which differ, not only with respect to the sampling frequency with which they are constructed, using either daily or high-frequency data, but also according to time span and asset class. We compare the
full sample parameter estimates and out-of-sample forecasting performance of our RLS-ARFIMA model
to six popular models in the literature and uncover some novel empirical findings. First, the random
level shift component is important for all series, delivering more frequent shifts for all volatility proxies
constructed from high-frequency data, but with less variability for most compared to those associated
with the daily return series. Second, once level shifts are taken into account, the high-frequency volatility
measures are characterized by a large genuine long memory component, whereas the remaining dynamics
of the volatility proxies constructed as log-absolute returns may be described as a combination of short
memory dynamics and measurement errors. Third, we show that if one fails to take both genuine
long memory and random level shifts into account, the resulting parameter estimates will reflect either
spurious long memory or spurious breaks. Most importantly, however, from our out-of-sample forecasting
analysis, we show that the RLS-ARFIMA model is, by far, the most frequent member of the 10% Model
Confidence Set proposed by Hansen, Lunde & Nason (2011). It consistently delivers good out-of-sample
performance across various forecast periods, forecast horizons, asset classes, and volatility measures.
The forecast gains can be very pronounced, especially at longer horizons.

The outline of the paper is as follows. Section 2 introduces the model and some motivational evi-
dence. Section 3 describes the model in a state space framework and introduces the forecasting procedure.
Section 4 treats measurement errors. The simulation study is presented in Section 5, while Section 6 con-
siders the empirical analysis. Finally, Section 7 concludes. A supplementary web appendix, Varneskov

2 The Volatility Model: Motivation and Specification

We seek to describe the dynamics of daily volatility measures for a variety of assets with observations
sampled at different frequencies. Hence, we first formulate a discrete time model for asset return volatility,
introduce the empirical volatility proxies, and provide some preliminary summary statistics.

2.1 A Discrete Time Volatility Model

As we aim to provide a unified discrete time framework for capturing the dynamics of daily volatility
measures, constructed from either daily- or high-frequency data, we need to specify a general time
series model that not only accommodates some of the extensively documented empirical regularities of
such processes such as, e.g., volatility clustering, genuine long memory and/or random level shifts, but
also allows for measurement errors in the volatility proxies. Let $x_t \in \mathbb{R}$ denote the latent, univariate
logarithmic volatility process, then we assume that the observable log-volatility proxy, $y_t \in \mathbb{R}$, behaves
according to the signal-plus-noise model:

$$x_t = a + h_t + v_t, \quad \text{(2)}$$

$$y_t = x_t + u_t \quad \text{where} \quad u_t \sim \text{i.i.d.} N(0, \sigma_u^2) \quad \text{(3)}$$
is a measurement error, $a$ is a constant, $h_t$ is a stationary long memory process, and $v_t$ is the random level shift component. In particular, we assume that the random level shift process is specified as $v_t = \sum_{j=1}^{T} \delta_{T,j}$ where $\delta_{T,j} = \pi_{T,j} \eta_j$ consists of a level shift of magnitude $\eta_j \sim i.i.d. N(0, \sigma_n^2)$ occurring with probability $\pi_{T,j} \sim i.i.d. \text{Bernoulli}(\gamma/T)$, for some $\gamma \in [0, T]$. The long memory component, $h_t$, is assumed to follow an ARFIMA process of the form $\Phi(L)(1-L)^d h_t = \Theta(L)e_t$ where $\Phi(L) = (1-\phi_1 L - ... - \phi_p L^p)$ and $\Theta(L) = (1 - \theta_1 L - ... - \theta_q L^q)$ are autoregressive and moving average lag $(Lh_t = h_{t-1})$ polynomials of orders $p$ and $q$, respectively, and $e_t \sim i.i.d. N(0, \sigma^2_e)$. Stationarity and identifiability are assured by letting $0 \leq d < 0.5$ and assuming that the roots of $\Phi(x)=0$ and $\Theta(x) = 0$ are outside the unit circle and distinct. Last, we assume that the components $\pi_{T,t}$, $\eta_t$, $u_t$ and $h_t$ are mutually independent.

Before we proceed, several features of the model should be highlighted. First, by imposing either $\gamma = 0$ or $\sigma_\eta = 0$, we recover the long memory stochastic volatility (LMSV) model and, if $\sigma_a = 0$ is additionally imposed, the stationary ARFIMA model advanced by Deo et al. (2006) and Andersen et al. (2003) in the context of realized volatility modeling and forecasting. This implies that if either $\gamma = 0$ or $\sigma_\eta = 0$, the other parameter affecting the random level shift process is not identified. This feature is evident in our simulation study in Section 5. However, and as we will elaborate upon in later sections, the likelihood function for the ARFIMA parameters are unaffected by this boundary case. Also, since we find both $\gamma > 0$ and $\sigma_\eta > 0$ for all series considered, and the main emphasis is on forecasting, the possibility of non-identified parameters is innocuous for the present analysis.

Second, if we impose $d = 0$, we recover a short memory stochastic volatility model with ARMA dynamics and random level shifts in the mean. We note that even this restricted version of the model generalizes the corresponding model in Qu & Perron (2013) by allowing for an MA component, and Lu & Perron (2010) by accommodating both an MA component and measurement errors in the series. Hence, our framework in (2)-(3) offers substantial flexibility when modeling the dynamics of various daily log-volatility measures. In particular, it allows us to remain agnostic as to whether the persistent features of the series are better described by genuine long memory, random level shifts, or both, and it may be applied to daily as well as high-frequency measures of volatility.

Third, we impose normality on $e_t$ and $u_t$, which may be restrictive considering that measurement errors for daily volatility proxies, in particular, can be highly non-Gaussian. The assumption, however, should be interpreted in a quasi-maximum likelihood (QML) sense. That is, we use it to derive the predictive likelihood function via the Kalman filter to estimate different versions of the model, similar to the strategy developed by, e.g., Harvey & Shephard (1996) for short memory stochastic volatility models, who show that consistency and asymptotic normality still hold when the measurement errors deviate from Gaussianity for a related QML estimator based on the Kalman filter. However, as we analyze logarithmic transformations of the volatility proxies, we do not expect to see dramatic violations of Gaussianity, cf.

\footnote{In the supplementary appendix, Varneskov & Perron (2015), we briefly discuss how the volatility in discrete time return models relate to the quadratic variation from continuous time return models. Moreover, we make a direct comparison of the \textit{discrete} signal-plus-noise model in (2)-(3) to a contemporaneous \textit{continuous time} stochastic volatility model.}

\footnote{In subsequent work, Grassi & de Magistris (2014) study the small sample properties of fractional integration order parameter estimates, $d$, using a simplified version of the proposed model (2)-(3) in a simulation setup.}
the distributional results in Andersen, Bollerslev, Diebold & Ebens (2001) and Andersen, Bollerslev, Diebold & Labys (2001).

Fourth, the accommodation of measurement errors in the signal-plus-noise model has implications for the reduced form dynamics of the observable log-volatility proxy, $y_t$. In particular, and similar to the analyses in Meddahi (2003) and Hansen & Lunde (2014), who assume that realized volatility proxies obey ARMA dynamics, we may reformulate the model as

$$(1 - L)^d \Phi(L)(y_t - a - v_t) = \Theta(L)\epsilon_t + (1 - L)^d \Phi(L)u_t. \tag{4}$$

This representation has implications for how we treat measurement errors and interpret the estimated MA parameters. A detailed discussion of these issues is deferred to Section 4.

Finally, we stress that the Bernoulli probability of a random level shift is dependent on the sample size, $T$, to make the expected number of shifts constant and equal to $\gamma$. This is needed to model structural changes in mean (or rare events), which affect the properties of the series until the next shift (event) occurs. The long memory component allows the process to have transitory shocks that are long-lasting in periods between structural changes. For example, in the context of volatility modeling, this may potentially capture volatility clustering between financial crises (which may be seen as rare events). If only one persistent component is present in the log-volatility series, our model is able to assess whether it is better described by genuine long memory or random level shifts.

**Remark 1.** The local level model $y_t = x_t + u_t$, $x_t = x_{t-1} + \epsilon_t$ put forth in, e.g., Harvey (1989), is embedded in our framework by imposing a level shift in each period, i.e., $\gamma = T$. However, in general, we require $\gamma \in (0, T)$ fixed such that $\gamma/T \to 0$ as $T \to \infty$ for the level shift component to generate autocorrelations akin to genuine long memory, see, e.g., Perron & Qu (2010). Furthermore, we find that $\gamma = T$ is strongly rejected for all series in our empirical analysis.

### 2.2 The Data and Construction of the Volatility Series

We consider eight daily log-volatility series in our empirical analysis, which differ, not only according to the sampling frequency of the data with which they are constructed, but also according to time span and asset class: (1) For three stocks, Bank of America Corp. (BAC), Merck & Co., Inc. (MRK), and the Standard & Poor’s Depository Receipts (SPY), we have tick-by-tick trades available with observations stamped to the nearest second from January 1997 through July 2008; (2) For futures contracts on the S&P 500 and 10-year Treasury bonds, we have one-minute observations available for every trading day from January 1983 through May 2009; (3) For the three exchange rates, USD-AUD, USD-CHF, and USD-YEN, we have daily observations available from January 4th 1971 through April 10th 2009.\(^4\)

The number of trading days, hence the time span, is considerably smaller for the volatility measures constructed from intra-daily data than for the daily volatility proxies. However, from the theory of quadratic variation, it is well-known that, under mild conditions on the efficient price process, we

\(^4\)We are grateful to Asger Lunde for providing cleaned tick data.
may utilize high-frequency data to get a precise estimate of the whole return variance trajectory over a (trading) day. In particular, if the applied estimator is able to account for an array of market frictions that are inherent to observable intra-daily log-prices, then high-frequency data-based estimates of quadratic variation make unbiased and efficient proxies, thus having measurement errors that are vanishingly small, which has been shown to improve out-of-sample forecasting in, e.g., Andersen et al. (2003), Koopman et al. (2005), Deo et al. (2006), and Varneskov & Voev (2013).

The volatility for the three daily exchange rate series is proxied by log-absolute returns.\(^5\) The daily quadratic variation, on the other hand, for the remaining series with high-frequency data available is estimated using the flat-top realized kernel approach put forth in Varneskov (2014, 2015) since it is robust to general forms of market microstructure noise and has optimal asymptotic- as well as good finite sample properties.\(^6\) Each flat-top realized kernel estimate is subsequently square-root- and log-transformed such that its unit is comparable to that of log-absolute returns. We provide a few unconditional- and conditional summary statistics of the eight volatility proxies in Table 1. From the unconditional summary statistics, we see that the three exchange rate volatility series exhibit slightly more left-skewed distributions with slightly higher excess kurtosis compared to the remaining series based on high-frequency data. However, it is clear that the logarithmic transformation has removed the pronounced right-skew and excess kurtosis, which usually characterize volatility proxies in their standard deviation or variance form. These distributional results are in line with prior findings in, e.g., Andersen, Bollerslev, Diebold & Ebens (2001) and Andersen, Bollerslev, Diebold & Lahys (2001).

As a gauge of the conditional properties of the series, we present log-periodogram (LP) and local Whittle (LW) estimates of the fractional integration order using a bandwidth \(\lceil T^{1/2} \rceil\). Furthermore, we include the results from the testing procedure by Perron & Qu (2010) of the null hypothesis that the volatility series have genuine long memory against the alternative of being comprised of level shifts and short memory dynamics, and a similar test by Qu (2011), which shares the same null hypothesis, but also allows the alternative to be a combination of genuine long memory and level shifts.\(^7\) The point estimates from the LP and LW estimators suggest that all volatility series are fractionally integrated with \(d > 1/2\), that is, within the non-stationary range. A feature to be discussed and explained later. However, from the two tests of whether the persistence is generated (exclusively) by a genuine long memory component, we find clear evidence against it for the USD-AUD and USD-JPY series, no evidence against it for the MRK and SPY series, and mixed evidence against it for the remaining series. This suggest that incorporating both a genuine long memory- and a random level shift component may be important for capturing the low-frequency variation in daily volatility series and, subsequently, for generating competitive volatility forecasts.

\(^5\)Strictly speaking, we use \(\ln(|r_t| + 0.001)\), \(r_t\) being the daily log-return, to bound zero daily returns away from minus infinity. This follows Lu & Perron (2010), Perron & Qu (2010), and Qu & Perron (2013).
\(^6\)We provide details on the flat-top realized kernel estimator and its implementation in the supplementary appendix.
\(^7\)We detail the testing procedures and the LP and LW estimators in the supplementary appendix where we also provide a more in-depth analysis of the conditional properties of the volatility series. This includes theoretical- and empirical results on the autocorrelation function for time series with genuine long memory, random level shifts, and measurement errors.
3 Econometric Methodology

In this section, we re-cast the reduced form model (4) in state space form to provide a feasible estimation and forecasting framework, generalizing the methodology in Perron & Wada (2009) and Lu & Perron (2010) by allowing for genuine long memory. Additionally, we provide an easily implementable forecasting procedure, which may also be used for previously proposed short memory-style random level shift models. In this section, we first treat the observable log-volatility process, \( y_t \), as having an MA component of (finite) order \( q \) and defer a detailed discussion of measurement errors to the next section.

3.1 State Space Representation

First, redefine the random level shift component, \( v_t \), as a random walk with innovations that obey a mixture of two normally distributed processes as follows

\[
v_t = v_{t-1} + \delta_{T,t} \quad \text{where} \quad \delta_{T,t} = \pi_{T,t} \eta_1 + (1 - \pi_{T,t}) \eta_0 \]

and \( \eta_{jt} \sim \text{i.i.d.} N(0, \sigma^2_{\eta_j}) \) for \( j = (0,1) \). We impose the restrictions \( \sigma^2_{\eta_1} = \sigma^2_{\eta_0} \) and \( \sigma^2_{\eta_0} = 0 \) to recover the representation in (2). The intuition for reducing the two components of the model to one is that if a structural change occur, it will have a long-lasting impact on the volatility level, at least until the next structural change. However, writing \( v_t \) using this “two-component-form” allows to adopt a state space representation that resembles the corresponding one for Markov regime switching models see, e.g., Hamilton (1994b), which is helpful in developing the estimation procedure. Moreover, this specification also highlights that level shifts are modeled as random events, which are invariant to past realizations of the data. Next, under the conditions of Section 2.1, the long memory component, \( h_t \), in (4) may be written as an AR(\( \infty \)) process,

\[
h_t = \sum_{i=1}^{\infty} \psi_i h_{t-i} + \epsilon_t, \quad \text{where} \quad \sum_{i=0}^{\infty} \psi_i L^i = \frac{\Phi(L)}{\Theta(L)} (1 - L)^d. \tag{5}
\]

Since \( 0 \leq d < 1/2 \), and the roots of \( \Phi(x)=0 \) and \( \Theta(x) = 0 \) are outside the unit circle and distinct, \( h_t \) has a unique and stationary solution, see Brockwell & Davis (1991, p. 525). The contribution of the fractional difference filter may be written as a binomial expansion \( (1 - L)^d = \sum_{i=0}^{\infty} \pi_i L^i \) with \( \pi_i = \Gamma(i - d)/(\Gamma(i + 1)\Gamma(-d)) \) where \( \Gamma(\cdot) \) is the gamma function. Using this representation, we may rewrite \( y_t \) in first differences as \( \Delta y_t = h_t - h_{t-1} + \delta_{T,t} \) for \( t = 2, \ldots, T \). Similar to the frameworks for ARFIMA models in Chan & Palma (1998) and Beran (1995), \( \Delta y_t \) does not have an exact finite dimensional state space representation unless \( d = 0 \) and \( p, q < \infty \). Hence, we follow the literature and approximate the AR(\( \infty \)) process by an AR(\( M \)) where \( M \) is chosen depending sample size (details are in Section 5). Hence, the approximate state space representation of (4) in matrix form is

\[
\Delta y_t = FH_t + \delta_{T,t}, \quad H_t = GH_{t-1} + E_t \tag{6}
\]
where \( F = (1,-1,0,\ldots,0)' \), \( H_t = (h_t,h_{t-1},\ldots,h_{t-M+1}) \), and \( E_t = (\epsilon_t,0,\ldots,0) \) are \( M \times 1 \) vectors, \( E_t \sim \text{i.i.d.} \mathcal{N}(0_{M \times 1}, Q) \) and \( 0_{M \times 1} \) denotes a \( M \times 1 \) vector of zeros. Here, \( G \) and \( Q \) are both \( M \times M \) matrices of parameters and identifying terms,

\[
G = \begin{pmatrix}
\Psi_{M-1} & \psi_M \\
I_{M-1} & 0_{(M-1) \times 1}
\end{pmatrix}, \quad Q = \begin{pmatrix}
\sigma_t^2 & 0_{1 \times (M-1)} \\
0_{(M-1) \times 1} & 0_{(M-1) \times (M-1)}
\end{pmatrix},
\]

where \( \Psi_M = (\psi_1,\ldots,\psi_M) \) is \( 1 \times M \) and \( I_M \) is a \( M \)-dimensional identity matrix. The added challenge relative to the genuine long memory state space framework of Chan & Palma (1998) is due to the state-dependent error in the measurement equation, whereas relative to Lu & Perron (2010) it is the presence of \((1 - L)^d/\Theta(L)\) in the representation of \( h_t \) such that no finite state space representation exists.

### 3.2 Maximum Likelihood Estimation

The basic principle behind the estimation procedure is to augment the probability of states by the realizations of a mixture of normally distributed processes at time \( t \) and apply the Kalman filter to construct the likelihood function conditional on the realization of states. Since we truncate the AR(\( \infty \)) representation of \( h_t \) in (5) at lag \( M \), the resulting estimation method becomes highly similar to the corresponding procedures in Perron & Wada (2009) and Lu & Perron (2010). Hence, details on the construction of the log-likelihood function are deferred to the supplementary appendix.

Note that if either \( \gamma = 0 \) or \( \sigma_\eta = 0 \), the other parameter is not identified and the estimation procedure collapses to the genuine long memory state space framework of Chan & Palma (1998). Defining the parameter vector \( \Sigma = (\sigma_\gamma,\gamma,\sigma_\epsilon,d,\phi_1,\ldots,\phi_p,\theta_1,\ldots,\theta_q)' \), then from their Theorems 3.1 and 3.2, the estimator of the ARFIMA parameters \( \Pi = \Sigma \setminus \{\gamma,\sigma_\eta\} \), denoted \( \hat{\Pi} \), is consistent when \( M = T^\beta \) with \( \beta > 0 \), and when \( \beta \geq 1/2 \), \( \sqrt{T}(\hat{\Pi} - \Pi) \overset{D}{\to} \mathcal{N}(0,\Lambda^{-1}(\Pi)) \) where \( \Lambda(\Pi) \) is the usual information matrix. In other words, the ARFIMA parameter estimates have the usual maximum likelihood properties and are unaffected by the possible event of non-identification of the random level shift parameters.

### 3.3 Forecasting with the RLS-ARFIMA Model

First, define \( Y_t = (\Delta y_2,\Delta y_3,\ldots,\Delta y_T)' \) and denote the filtered state vector by \( H_{i|t}^{ij} \) along with its associated covariance matrix by \( P_{i|t}^{ij} \), both of which depend on whether \( \pi_{T,t-1} = i \) and \( \pi_{T,t} = j \) for \( i,j \in \{0,1\}^2 \), that is, on whether a random level shift occurs at either time \( t-1 \), time \( t \), or both. Then, the state space structure of the RLS-ARFIMA(\( p,d,q \)) model allows us to obtain \( \tau \)-step-ahead forecasts by utilizing the model formulation and combining results from the state space and Markov regime switching forecasting literature, see, e.g., Brockwell & Davis (1991), Hamilton (1994a), and Gabriel & Martins (2004). Formally, we forecast according to the following proposition:

**Proposition 1.** Let \( y_t \) satisfy the conditions of Section 2.1 and let \( \mathbb{E}_t[y_{t+\tau}] = \hat{y}_{t+\tau|t} \) denote the expected value of the process at time \( t+\tau \), conditional on the information available at time \( t \), then the \( \tau \)-step-ahead
forecast is

\[ \hat{y}_{t+\tau | t} = y_t + F G^T \sum_{i=0}^{1} \sum_{j=0}^{1} \Pr(\pi_{T,t+1} = j) \Pr(\pi_{T,t} = i | \pi_{T,t}) \frac{H_{ij}}{\sum_{l=0}^{1} \sum_{k=0}^{1} \Pr(\pi_{T,t+1} = l) \Pr(\pi_{T,t} = k | \pi_{T,t})}. \]

**Proof.** See the supplementary appendix, Varneskov & Perron (2015).

Proposition 1 illustrates two corrections relative to standard ARFIMA forecasts, e.g., Brockwell & Davis (1991) and Doornik & Ooms (2004). The first is a mean correction where, in particular, the unconditional mean is replaced by \( y_t \). Intuitively, if the underlying process contains random level shifts, an unconditional mean has no information as to which regime the process is in at time \( t \), but this is reflected in \( y_t \). The second is a path correction. A \( \tau \)-step-ahead forecast for each realization of the state is made, \( G^T \frac{H_{ij}}{\sum_{l=0}^{1} \sum_{k=0}^{1} \Pr(\pi_{T,t+1} = l) \Pr(\pi_{T,t} = k | \pi_{T,t})} \), and then weighted by the probability of being on a given transition path between regimes, \( \Pr(\pi_{T,t+1} = j) \Pr(\pi_{T,t} = i | \pi_{T,t}) \). As the second component of the forecast in Proposition 1 is transitory, we have \( \hat{y}_{t+\tau | t} \to y_t \) as \( \tau \to \infty \). This implies that the forecasts obtained from the short memory random level shift model of Lu & Perron (2010) and from our RLS-ARFIMA(\( p, d, q \)) model coincide for long horizons, but differ for short to intermediate horizons.

**Remark 2.** The proposed forecasting framework encompasses multiple types of forecasting schemes; recursive estimation using an expanding window of observations, rolling window of observations, and a one-time estimation of the parameters, which, in conjunction with the Kalman recursions, may be used to generate forecasts conditional on the parameter estimates.

### 4 Handling Measurement Errors

The econometric methodology relies on an RLS-ARFIMA(\( p, d, q \)) approximation of the reduced form dynamics, that is, a finite-order ARMA representation to account for both the underlying short memory dependencies and measurement errors in the series. We discuss the validity of our approach and provide evidence in favor of specific parameterizations.

#### 4.1 Measurement Errors and the ARFIMA representation

From the reduced form of \( y_t \) in (4), we immediately see that by allowing for measurement errors in the observable log-volatility proxy, we generally need an RLS-ARFIMA(\( p, d, \infty \)) structure to fully capture its dynamics. However, if \( d = 0 \), then (4) illustrates that we may approximate the residual dynamics, i.e., the dynamics once level shifts are taken into account, by a finite ARMA(\( p, \max(p, q) \)) specification, see Granger & Morris (1976). If, on the other hand, \( d > 0 \) and we have no measurement errors, then an ARFIMA(\( p, d, q \)) model will fully capture the residual dynamics. Our evidence, presented in Section 2 and below, and the later empirical analysis suggest that these two cases describe the volatility proxies constructed from daily, respectively, high-frequency data, and that an AR(1) and ARMA(1,1) specification adequately capture the short memory dependencies of the former, respectively, the latter.
A model that encompasses both these cases is the RLS-ARFIMA(1, d, 1) specification, which we analyze in detail later in the simulation study and the empirical analysis.

4.2 Motivational Evidence from an RLS-LMSV(1, d) Model

We directly estimate the parameters of the model (2)-(3) using a filtered long memory stochastic volatility model with random level shifts, or the RLS-LMSV model. This extends the discrete-time SV models in, e.g., Deo et al. (2006) and Qu & Perron (2013). To estimate said model, we propose the following modifications to the state space model described in Section 3.

Let the observable log-volatility proxy, $y_t = x_t + u_t$, be written in truncated state space form as

$$y_t = \tilde{F}\tilde{H}_t + u_t, \quad \text{with} \quad \tilde{H}_t = \tilde{G}\tilde{H}_{t-1} + \tilde{T}\tilde{E}_{t,\pi}$$

where $\tilde{F} = (1, 0, \ldots, 0, 1)'$ and $\tilde{H}_t = (H_t, v_t)$ are $(M + 1) \times 1$ vectors, and the $2 \times 1$ vector $\tilde{E}_{t,\pi} \sim \text{i.i.d. } N(0_{2 \times 1}, \tilde{Q}_\pi)$ depends on the particular regime of the process at time $t$ through

$$\tilde{Q}_1 = \begin{pmatrix} \sigma_e^2 & 0 \\ 0 & \sigma_\eta^2 \end{pmatrix} \quad \text{and} \quad \tilde{Q}_0 = \begin{pmatrix} \sigma_e^2 & 0 \\ 0 & 0 \end{pmatrix}.$$

Furthermore, by defining the $(M + 1) \times (M + 1)$ and $(M + 1) \times 2$ matrices

$$\tilde{G} = \begin{pmatrix} G & 0_{M \times 1} \\ 0_{1 \times M} & 1 \end{pmatrix}, \quad \text{and} \quad \tilde{T} = \begin{pmatrix} 1 & 0_{M \times 1} \\ 0_{M \times 1} & 1 \end{pmatrix},$$

we may apply an estimation procedure similar to the one described in the supplementary appendix.

Before proceeding to the estimation results, we assess the accuracy of the RLS-LMSV parameter estimates from the proposed methodology by simulating an RLS-LMSV(1, d) process with $d = 0.35$, $\gamma/T = 0.02$, $\sigma_e = 0.5$, $\sigma_\eta = 3\sigma_e$, $\phi = 0.2$ and two different levels of measurement errors specified through the noise-to-signal ratio, $\xi = \sigma_u^2/\sigma_e^2(1-\phi)^2$ with $\xi = \{1, 2\}$. We compute the bias and root mean squared error (RMSE) of the estimates for sample sizes $T = \{3000, 6000\}$, truncations $M = \{20, 30, T^{1/2}\}$ of the AR($M$) representation, and $N = 100$ replications. The results are presented in Table 2. We defer a discussion of the specific setup, i.e., truncation, implementation, etc., to our main simulation study in the next section. For now, we are mainly interested in whether we can identify the key parameters $\gamma/T$, $\sigma_\eta$, $d$, and $\sigma_e$ for fairly high levels of measurement noise. The latter is chosen higher than what our empirical estimates suggest, except for the USD-YEN series, to provide a conservative assessment.

Table 2 illustrates two important points. First, we observe that the RLS-LMSV(1, d) model consistently recovers the random level shift parameters $\gamma/T$ and $\sigma_\eta$ with no or a vanishingly small bias. Second, the estimates of the genuine long memory parameter, $d$, is, not surprisingly, slightly downward biased. However, this bias is diminishing when increasing the sample size, truncation length, or decreasing the noise-to-signal ratio (as stressed above, the bias is to be interpreted as a conservative upper
bound). Then, to provide support for our RLS-ARFIMA models in Section 4.1, we report the parameter estimates from an RLS-LMSV(1, d) model for all eight log-volatility series in Table 3 using $M = 20$.

From Table 3, we make a few noteworthy observations. First, for the three log-volatility series constructed from tick-by-tick trades, BAC, MRK, and SPY, we see that the impact of measurement errors is negligible, and similarly for the S&P 500 series. For the three exchange rate series, on the other hand, we observe non-negligible measurement noise. However, the estimated noise-to-signal ratios for the USD-AUD and USD-CHF series are still (much) smaller than the corresponding simulated values, but similar for the USD-YEN series. Moreover, note that we hardly estimate any ARFIMA dynamics for the former two. In this case, we cannot separately identify $\sigma_\epsilon$ and $\sigma_u$ since the parameter will collectively measure the noise level in the series. As we will see in the next section, if we simply interpret the noise as coming from one source, here $\sigma_\epsilon$, the RLS-ARFIMA model precisely recovers this parameter.

In general, the estimated parameters in Table 3 present a striking pattern across the volatility series. Random level shifts are present in all series, occurring more frequently for all volatility proxies constructed from high-frequency data, but with less variability for most compared to those associated with the daily return series. In addition, the high-frequency volatility measures contain a large genuine long memory component whereas there are seemingly little ARFIMA dynamics remaining in the exchange rate volatility series once level shifts have been accounted for. Finally, we also observe a combination of measurement errors, random level shifts, and genuine long memory for the high-frequency-based T-bond series. However, given equivalent representations of AR(1) plus noise and ARMA(1, 1) dynamics and our later empirical findings for the RLS-ARFIMA model, we cannot exclude that the former is caused by a negative MA(1) component, which appears prominently in our later analysis.

These results highlight the relevance of the RLS-ARFIMA modeling strategy proposed in Section 4.1, on which we will focus in the remainder of the paper.

5 Simulation Study

In this section, we investigate the accuracy of the parameter estimates from the state space estimation methodology. To show the validity of our proposed estimation method, and to get an indication of how to select $M$, the order of truncation of the AR($M$) representation, we set up a simulation study focused to distinguish the proportion of persistence in the time series attributed to random level shifts and to genuine long memory. Additionally, we compare the resulting parameter estimates to ones obtained from fitting ARFIMA($p, d, q$) models to gauge how these are affected by level shifts, and, finally, how our estimates of the (RLS-)ARFIMA($p, d, q$) parameters compare when level shifts are absent, since the presence of two non-identified parameters may lead to efficiency losses.

We consider a Monte-Carlo study with $N = 100$ replications, four different truncation lengths $M = \{5, 10, 20, T^{1/2}\}$, and a sample size $T = 3000$.\footnote{We also performed some simulations for sample sizes $T = 1000$ and $T = 5000$, which showed proportionally worse/better results. Ideally, we would like to carry out the simulations for $N \gg 100$. However, this presents a computation challenge, in particular for large $M$. Hence, the results should be interpreted as indicative rather than definitive.} The data generating processes (DGP's) are simulated
from an RLS-ARFIMA(1, d, 1) model with $\phi = 0.2$, $\theta = -0.1$, $\sigma_\epsilon = 0.5$ and (DGP 1) $d = 0$, $\gamma/T = 0.02$, $\sigma_\eta = 3\sigma_\epsilon$, (DGP 2) $d = 0.35$, $\gamma/T = 0$, $\sigma_\eta = 0$, (DGP 3) $d = 0.35$, $\gamma/T = 0.02$, $\sigma_\eta = 3\sigma_\epsilon$, and (DGP 4) $d = 0.6$, $\gamma/T = 0.02$, $\sigma_\eta = 3\sigma_\epsilon$. The choices of $M$ are motivated by Chan & Palma (1998) and Martin & Wilkins (1999), who find that a small truncation order suffices to capture the dynamics of an ARFIMA process. The choice of the sample size is motivated by the typical length of financial time series. The selected parameters of the first three DGP’s are based on parameter estimates from the level shift literature for DGP 1, e.g., Qu & Perron (2013), from the long memory volatility modeling literature for DGP 2, e.g., Andersen et al. (2003), and from one that combines them for DGP 3, e.g., this paper’s estimates for the S&P 500 series. We include DGP 4 as a robustness check to ensure that our empirical detection of random level shifts is not spuriously caused by a non-stationary fractionally integrated component. Finally, we have also fitted an RLS-ARFIMA(1, d, 1) model to the simulated RLS-LMSV(1, d) process from the previous section. The results for the key persistence parameters are highly similar to the ones reported in Table 2 and are, thus, omitted.

5.1 Implementation

For estimation of the RLS-ARFIMA models, we know that all components in the state vector (6) are stationary. Hence, we initialize the updating equations using their unconditional expected values, $H_{ij}^{0|0} = 0_{M \times 1}$ and $P_{ij}^{0|0} = Q$. To start the probability weighting of the likelihood function, we set $\Pr(\pi_{T,0} = 1|Y_0; \Sigma) = \gamma/T$. Lastly, we draw the initial values of the parameters from a uniform distribution five times and select the optimized estimates with the highest associated log-likelihood value.

The ARFIMA($p, d, q$) models in (5) are estimated using the conditional-sum-of-squares (CSS) estimator, c.f. Beran (1995) and Nielsen (2015), since the presence of random level shifts in the series may potentially induce an upward bias in the estimated integration order, and the CSS estimator is valid for stationary as well as non-stationary values of $d$. The residual standard deviation, $\sigma_\epsilon$, is estimated as $\hat{\sigma}_\epsilon = \sqrt{(T - 1)^{-1}\sum_{t=1}^{T} \hat{\epsilon}_t^2}$ where $\hat{\epsilon}_t$ is the model-implied residuals. For all (RLS-)ARFIMA models, we restrict attention to the (0, $d$, 0) and (1, $d$, 1) parameterizations, given the empirical results in the previous section and since simpler models are often advocated for out-of-sample forecasting.

5.2 Simulation Results

The bias and RMSE of the parameter estimates for all estimators and DGP’s are presented in Table 4. For DGP 1, we observe that the RLS-ARFIMA(1, d, 1) estimate of $\sigma_\eta$ is slightly upward biased, and that the model provides precise estimates of $d$ and $\gamma/T$. The relative difference in the estimate of $d$ obtained from the ARFIMA(1, d, 1) model, on the other hand, is quite suggestive, and while the evidence is provided in a stylized setup, we observe exactly the same pattern in our empirical analysis below. As documented by Perron & Qu (2010), if random level shifts are present in the series, the resulting estimate of $d$ obtained from an ARFIMA(1, d, 1) model will be inflated to capture the large estimates of $d$ obtained from a log-periodogram regression with few frequency ordinates. In order to
capture the smaller estimates when more frequency ordinates are included, the fitted MA parameter is biased towards a large negative value to accentuate the short-run mean reversion. Similarly, we find the ARFIMA(0, d, 0) estimate of d to be upward biased, yet the bias is not as dramatic as for the ARFIMA(1, d, 1) model since the former lacks an MA parameter to help fit movements at higher frequencies. Last, for the RLS-ARFIMA(0, d, 0) model, we see that the inclusion of positive short-run dynamics in the DGP causes d to be overestimated. This holds true for all DGP’s considered.

The results for DGP 2 verify that the ARFIMA(1, d, 1) model parameters are precisely estimated, as expected. What is particularly interesting for the present analysis, however, is that the ARFIMA parameters of the RLS-ARFIMA(1, d, 1) model are estimated with the same precision. As emphasized in Section 3.2, this may be explained by the fact that when \( \sigma_\eta \to 0 \) (which occurs when the truncation order, \( M \), increases), the estimation method collapses to the genuine long memory state space framework of Chan & Palma (1998). Consequently, the ARFIMA parameter estimates do not suffer from efficiency losses when the random level shift parameters are not identified.

For DGP 3, we observe that when the specification is tailored to the reduced form model, all the parameter estimates are unbiased and precise, while the corresponding estimates for the ARFIMA(1, d, 1) model display exactly the same bias as for DGP 1. Furthermore, the almost identical results for DGP 4 illustrates that the RLS-ARFIMA model does not confuse random level shifts with non-stationary fractional integration. As such, the proposed model is able to distinguish between the proportion of persistence attributed to random level shifts and genuine long memory. Since the bias of the various memory parameters are generally decreasing in \( M \), we select \( M = 20 \) as the order of truncation for the empirical analysis as there is a tradeoff with computational speed, especially for the longer series of daily returns. Finally, we increase the number of draws of the initial values to 10 to ensure that we do not report results from a local maximum.

**Remark 3.** The choice of truncation is important and, as a robustness check, we have experimented with selections \( M = \{ 30, 40 \} \) in both the simulation study and in the empirical analysis below. The results are almost identical to those obtained for \( M = 20 \). Similarly, the parameter estimates using the theoretically consistent choice \( M = \lfloor T^{1/2} \rfloor \) for the volatility measures based on high-frequency data are highly similar to those reported. This robustness to the choice of (a smaller) truncation order are in line with the findings in Chan & Palma (1998) and Martin & Wilkins (1999).

## 6 Empirical Analysis of Asset Return Volatility

We proceed demonstrating the relevance of the proposed reduced form (log-)volatility modeling and forecasting framework by comparing the full-sample parameter estimates and out-of-sample forecasting performance of specific RLS-ARFIMA models to other widely applied models in the discrete time volatility literature. Initially, we consider parameter estimates from the RLS-ARFIMA(0, d, 0), RLS-ARFIMA(1, d, 1), RLS-ARMA(0, 0), RLS-ARMA(1, 1), ARFIMA(0, d, 0), and ARFIMA(1, d, 1) models for three reasons. First, it allows us to assess whether the most persistent component in the series is
better described by random level shifts and/or genuine long memory and the impact of neglecting either one on the parameter estimates. Second, less parameterized models are often advocated for forecasting, see, e.g., Andersen et al. (2003). Third, as argued in Section 4, smaller order parameterizations suffice to capture both the short-run dynamics and measurement errors in the volatility measures.

In the forecasting exercise, we also include the six models mentioned above. The ARFIMA class of models has recently received much attention in the volatility prediction literature. For example, it has been shown in, among others, Andersen et al. (2003), Koopman et al. (2005), Deo et al. (2006), Chiriac & Voev (2011), and Varneskov & Voev (2013) to outperform the popular class of GARCH models in terms of out-of-sample forecasting when applied to high-frequency measures of volatility. Similarly, Lu & Perron (2010) and Qu & Perron (2013) find that short memory-style random level shift models provide forecasts, which are, at least, on par with those obtained from (FI)GARCH and discrete time SV models when applied to volatility proxies constructed from daily data. Hence, to examine the usefulness of the proposed RLS-ARFIMA model in different settings, we compare its out-of-sample forecasting performance to these state-of-the-art competitors. In addition, we include the HAR model introduced by Corsi (2009), which has been shown to provide accurate forecasts for realized volatility measures, and a benchmark GARCH(1, 1) model in our out-of-sample analysis.

Finally, note that we will describe the results for the SPY and USD-YEN series in details throughout since they represent two different groupings of the series (SPY: BAC, MRK, S&P 500) and (USD-YEN: USD-AUD, USD-CHF), which share similar characteristics within each group. The T-bond series, on the other hand, is harder to classify as it sometimes shares characteristics with the SPY group and sometimes with the USD-YEN group. We will make the distinction clear when necessary.

6.1 Full-Sample Parameter Estimates

We report the results for the eight log-volatility series in Tables 5-6. In particular, note that the results for the SPY series are presented in Panel C of Table 5, and those for the USD-YEN series in Panel D of Table 6. We first discuss the results for the SPY series. The estimated persistence parameters of the RLS-ARFIMA(0, d, 0) model are \( d = 0.4181 \) and \( \gamma/T = 0.0177 \), which suggests the joint presence of genuine long memory and random level shifts, similar to the results in Table 3. The estimated probability of level shifts indicates that they occur with an average duration of 56 days. Said duration is fairly low compared to the results in Lu & Perron (2010) for daily log-absolute returns on the S&P 500, AMEX, Dow Jones, and NASDAQ. We obtain similar estimates of the persistence parameters for the RLS-ARFIMA(1, d, 1) model in addition to large and significant estimates of the two ARMA parameters. The latter, however, seem to characterize a common factor, they have fairly high standard errors, and their inclusion hardly increases the log-likelihood value. This clearly suggests that the most important sources of variation is captured by the joint modeling of genuine long memory and random level shifts. The results from the ARFIMA(0, d, 0) model similarly indicate the presence of a stationary genuine long memory component, while the corresponding estimate for the ARFIMA(1, d, 1) model of \( d = 0.5965 \) suggests that the series

\(^{9}\text{The associated standard errors are computed using the (inverse) numerical Hessian matrix.}\)
is a non-stationary, fractionally integrated process. Furthermore, we observe that the estimated ARMA parameters of the ARFIMA(1, d, 1) model are large and distinct, however insignificant. As explained in the simulation study, this particular difference between the RLS-ARFIMA and ARFIMA parameter estimates is exactly what we expect when a random level shift component is present; the estimate of \( d \) is biased upwards to capture movements at the lower frequencies, while the MA parameter is biased towards a large negative value to accentuate the short-run mean reversion. When accounting for random level shifts, such biases are no longer present, and the genuine long-memory component is seen to be stationary with the remaining short-run variation close to being serially uncorrelated. Finally, the estimated probability of a random level shift using the RLS-ARMA(0, 0) and RLS-ARMA(1, 1) models are \( \gamma/T = 0.2082 \) and \( \gamma/T = 0.0797 \), respectively, suggesting that level shifts, which are assumed to be rare events, occur with very low durations. This is clearly empirical evidence of spurious breaks. That is, when a genuine long memory component is present in the log-volatility series, the RLS-ARMA models are attempting to fit the additional persistence by overestimating the number of shifts.

Next, consider the parameter estimates from the RLS-ARFIMA(0, d, 0) model for the USD-YEN series. The persistence parameters are \( d = 0.05 \) and \( \gamma/T = 0.0027 \), both statistically significant. The former, however, while deemed statistically significant indicates that the genuine long memory component is essentially irrelevant for characterizing persistent movements in the series. The estimated probability of random level shifts suggests that they are rare (26 in 9600 days) and occur with an average duration of 370 days. However, their magnitude \( \sigma_\eta = 3.0657 \), in comparison with the residual standard deviation, \( \sigma_\epsilon = 1.2765 \), demonstrates that they are large contributors to the total variation in the series. The results for the RLS-ARFIMA(1, d, 1) model are similar; the impact of random level shifts is almost identical, and the estimate of \( d \) is even smaller with a value of 0.00. Moreover, we observe that the ARMA coefficient estimates are both high and of similar magnitude, which is consistent with the interpretation in Section 4 that the daily log-volatility measure exhibits a combination of AR(1) residual dynamics and measurement errors. Unlike the results for the SPY series, the ARMA parameters for the USD-YEN series are seen to have small standard errors, and their inclusion increases the log-likelihood value, especially relative to the RLS-ARMA(0, 0) case. For the latter, we observe an estimated probability of random level shifts that is twice as high, which is, again, suggestive of positively dependent residual dynamics, though not as strong as for the SPY series.

Given the evidence from the RLS-ARFIMA models that random level shifts describe the low-frequency movements in the USD-YEN series, it is interesting to consider the estimated integration orders from the ARFIMA(0, d, 0) and ARFIMA(1, d, 1) models, which are, in contrast, but as expected, much higher and significant. Again, we observe interesting differences between the two models. The estimate of \( d \) is much higher for the ARFIMA(1, d, 1) model since it has a large negative MA component to induce strong mean reversion. These features are similar to the ones obtained for the SPY series, along with those in the simulation study, and they support the findings of random level shifts in the series.

The results are highly similar within each of the two groups, and the T-bond series seems to be better characterized by those obtained for the SPY group. Thus, we may draw some conclusions from
our analysis so far. The random level shift component is important for all series, being more frequent for all high-frequency measures of volatility, but with less variability for most. Once this is taken into account, the SPY group still contains a large, genuine long memory component. The remaining dynamics for the USD-YEN group, on the other hand, may be described using a positively dependent short memory component in combination with measurement errors. The theoretical causes of this puzzling difference between the reduced form dynamics of the return volatility series constructed from daily and high-frequency data is an interesting subject for further investigation.

6.2 Forecasting Performance Evaluation

The class of RLS-ARFIMA models allow for a more flexible description of the low-frequency variation in log-volatility series. However, whether such flexibility enhances out-of-sample volatility forecasts remains to be determined. Hence, we investigate the usefulness of the RLS-ARFIMA approach by comparing its forecasting performance to that from each of the competing dynamic models presented earlier along with the HAR model and the GARCH(1,1) benchmark, whose specifications and implementation procedures are briefly described in the supplementary appendix. This section proceeds by laying out the forecast evaluation framework before presenting the results from the out-of-sample exercise.

6.2.1 Forecast Evaluation Framework

We consider out-of-sample forecasting over the last $T_{out} = 900$ days. The various model parameters are estimated once, without the last 900 days in the sample, and the forecasts are computed conditional on these estimates. As we seek to evaluate the performance of direct $\tau$-step-ahead forecasts for three different horizons, $\tau = \{1, 5, 10\}$, let the cumulative forecast be defined as $y_{t+\tau,i|t} = \sum_{s=1}^{\tau} y_{t+s,i|t}$ for model $i \in M^0$ where $M^0$ is the initial finite set of models and, similarly, let the cumulative log-volatility proxy be denoted by $\hat{\sigma}_{t,\tau} = \sum_{s=1}^{\tau} y_{t+s}$. Then, we apply the mean squared forecast error (MSFE) criterion for the out-of-sample evaluation, $\text{MSFE}_{\tau,i} = \frac{1}{T_{out}} \sum_{t=1}^{T_{out}} \left( \hat{\sigma}_{t,\tau} - y_{t+\tau,i|t} \right)^2$, which was shown by Hansen & Lunde (2006) and Patton (2011) to be robust against measurement errors in the log-volatility proxy. To facilitate model comparison, define the relative performance of models $i, j \in M^0$ at time $t$ as $d_{ij,t} = \left( \hat{\sigma}_{t,\tau} - y_{t+\tau,i|t} \right)^2 - \left( \hat{\sigma}_{t,\tau} - y_{t+\tau,j|t} \right)^2$, for which, we assume the sequence $(d_{ij,t})$, $\forall i, j \in M^0$, $t = 1, \ldots, T_{out}$ satisfies the following conditions: For some $r > 2$ and $\gamma > 0$, $E[|d_{ij,t}|^{r+\gamma}] < \infty$, and $(d_{ij,t})$ is strictly stationary with variance $\text{Var}[d_{ij,t}] > 0$ and $\alpha$-mixing of order $-r/(r - 2)$.

Remark 4. These conditions impose restrictions on the sequences of relative forecast performances, $(d_{ij,t})$, not directly on the loss function, which is allowed to exhibit structural breaks, genuine long memory, etc. They seem to be satisfied by plots of the loss differentials and the robustness of our results to the use of recursive and rolling estimation windows. Even in the event that the conditions for

---

10This approach is chosen due to the heavy computational task of re-estimating parameters in each step for the group of (RLS-)ARFIMA models. As robustness checks, however, both recursive and rolling window estimation procedures have been used for some of the series; the numerical results are similar, and the model rankings are identical. This is explained by the parameter estimates being fairly robust to the choice of the estimation window.
the validity of the MCS evaluation procedure are violated, the numerical MSFE’s will provide a strong indication of the relative model performance.

Under the stated conditions on the sequence of loss differentials, we may assess the relative forecast accuracy of the models using the 10% Model Confidence Set (MCS) of Hansen et al. (2011), see the supplementary appendix for a review. It is important for our application that the MCS is based on a bootstrap implementation, which is robust against comparisons of nested models when the parameters are estimated once using the same in-sample period for all models, see, e.g., the discussions in Giacomini & White (2006) and Hansen et al. (2011). The MSFE’s and accompanying MCS p-values (in parentheses) are reported in Tables 7-9, where we use boldface notation to indicate whether a model belongs to the 10% MCS. As a robustness check, we have decomposed the out-of-sample results for the SPY and USD-YEN series into three non-overlapping sub-periods in Tables 7 and 8, respectively.

6.2.2 Out-of-Sample Results

First, to assist interpretation of the results, we illustrate how to read Tables 7-9 by considering the relative forecasting performance of the HAR model over the whole out-of-sample period of 900 days for the SPY series, which is reported in the bottom-right panel of Table 7. In this case, we observe that the HAR model belongs to the 10% MCS for one-step-ahead predictions, but not for five- nor ten-step-ahead forecasts, which implies that the model is significantly worse than the best set of dynamic models at predicting log-volatility for horizons of five and ten days.

In general, we find that when considering the SPY series and the whole out-of-sample period it is only the RLS-ARFIMA(1, d, 1) model that belongs to the 10% MCS for all forecast horizons, thus ranking as the best overall model. The RLS-ARFIMA(0, d, 0) ranks as the second best in terms of numerical MSFE’s. The RLS-ARMA(0, 0) and RLS-ARMA(1, 1) models also perform well for longer horizons, whereas the ARFIMA(0, d, 0), ARFIMA(1, d, 1) and HAR models do well for one-step-ahead predictions, but display MSFE’s of, at least, a factor three larger for ten-step-ahead predictions. This clearly shows the value of applying the proposed forecast procedure, which leads to significant gains in terms out-of-sample precision with the largest gains attributed to the mean correction. When the forecast performance is decomposed into three non-overlapping sub-periods, the RLS-ARFIMA(1, d, 1) model performs well in all cases, and its relative superiority over the remaining RLS-AR(FI)MA specifications is driven, in part, by the last 300 days of the sample. Note, however, that the forecast errors for the RLS-AR(FI)MA models are also the largest in this sub-period, while the discrepancy to the remaining models is the smallest, suggesting that, not surprisingly, it is difficult to pin down the mean of the series during the period covering the financial turmoil of late 2007 through July 2008. When the mean-behavior of the series is slightly less erratic, as during the first 600 out-of-sample days, the RLS-ARFIMA models performs much better than models that do not allow for random level shifts in the mean.

We proceed to evaluate the out-of-sample performance of the eight dynamic models using the USD-YEN series in Table 8 and immediately observe a similar model ranking; the RLS-ARFIMA(1, d, 1) model is significantly the best forecasting model for all horizons, followed by the RLS-ARFIMA(0, d, 0),
RLS-ARMA(0, 0) and RLS-ARMA(1, 1) models, which comprise a clear second tier.\textsuperscript{11} If we consider the evidence from Table 6 that the ARFIMA class of models display severely upward biased estimates of the (fractional) integration order, it is not surprising that we find these - along with the HAR and GARCH(1, 1) - models to exhibit much larger forecast errors, especially for longer horizons. In particular, this follows since they are not flexible enough to adequately describe the low-frequency variation in the volatility series and, thus, mistakenly summarizes the persistence as largely determined by a large genuine long memory component. Moreover, when decomposing the relative forecast performance into three non-overlapping samples, we see that all models, not surprisingly, deliver the largest forecast errors during the last 300 days, which cover most of the recent financial crises, and we observe that the RLS-ARFIMA(1, d, 1) model consistently exhibits the smallest MSFE's across sub-periods.

Finally, we may generalize the conclusions from the SPY and USD-YEN series by considering the out-of-sample results for the six remaining series in Table 9. Aggregating the results across the volatility series and forecast horizons, the RLS-ARFIMA class of models belong to the MCS in 21/24 cases, the RLS-ARMA class in 16/24 cases, the HAR model in 8/24 cases, the ARFIMA class in 7/24 cases, and the GARCH model never belongs to the MCS. Furthermore, we observe large reductions in the MSFE's with models that explicitly capture random level shifts. This clearly shows the importance of our proposed mean, or level shift, correction for out-of-sample forecasting, and it illustrates that the transitory path correction provides additional (albeit smaller) gains. The comparatively poor out-of-sample performance of dynamic models that do not explicitly model random level shifts, is, in itself, indirect evidence of their presence. As discussed previously, if level shifts are present, they bias the estimate of \(d\) upwards for the ARFIMA models (often in the non-stationary region) and the estimate of the MA parameter towards a large negative value. Similar biases affect the HAR and GARCH models, and they are responsible for the deterioration of the out-of-sample performance.

In general, we observe a good correspondence between in-sample fit and out-of-sample performance. The only exception arises if we contrast the parameter estimates from the RLS-AR(FI)MA models for the T-bond series, as shown in Table 6, with their respective out-of-sample results, where we see that the inclusion of ARMA parameters improves the in-sample fit, but leads to deteriorating out-of-sample performance as the forecast horizon increases. To elaborate on this observation, we depict the ten-step-ahead out-of-sample volatility for the T-bond series in Figure 1 together with the corresponding loss differentials from a bivariate comparison of the RLS-ARFIMA(0, \(d, 0\)) model against the RLS-ARFIMA(1, \(d, 1\)) model and a comparison of the RLS-ARMA(0, 0) model against the RLS-ARMA(1, 1) model. From the three series, we observe a distinct pattern; after an abrupt change around day 400, the log-volatility level is gradually increasing until, approximately, day 750. The less parameterized RLS-ARFIMA(0, \(d, 0\)) and RLS-ARMA(0, 0) models are better at capturing this increase, suggesting that the inclusion of ARMA parameters, in particular a strongly mean-reverting MA component, induces over-smoothing of the log-volatility series. This eventually leads to the deterioration in forecast performance.

\textsuperscript{11}The difference between the RLS-ARMA(1, 1) and RLS-ARFIMA(1, \(d, 1\)) models may, given the parameter estimates in Table 6, seem surprising. However, when we remove the last 900 days to avoid using in-sample information for estimation of the parameters, we observe minor differences between the parameter estimates from the two models.
as the mean-reverting log-volatility level deviates from its increasing out-of-sample counterpart for 350 observations. On the other hand, the inclusion of ARMA parameters seemingly enhances the forecast performance of the models during the first part of the sample. This suggests that further out-of-sample gains may potentially be extracted by constructing forecast combinations of the dynamics models. However, a deeper investigation of this potential is beyond the scope of the paper.

In summary, there is overwhelming evidence in favor of using the RLS-ARFIMA class of models, which is not only able to distinguish between the contributions from random level shifts and genuine long memory to the low-frequency variation of the log-volatility series, but also delivers consistently good out-of-sample performance across a variety of forecast periods, forecast horizons, asset classes, and volatility proxies with varying degrees of measurement errors.

7 Conclusion

We propose a reduced form framework for modeling the volatility of asset returns, which allows for the presence of random level shifts, genuine long memory and measurement errors. In particular, we advocate a parametric state space model where the underlying dynamics is decomposed into a simple level shift component and ARFIMA dynamics. This allows both long- and short memory parameters to be estimated together with the probability and magnitude of random level shifts. Measurement errors are accounted for by careful modeling and interpretation of the ARMA parameters. We provide an estimation procedure and a forecasting framework to construct mean- and path-corrected forecasts.

We perform an empirical analysis using eight daily return volatility series, which differ, not only according to the sampling frequency of the data with which they are constructed, but also with respect to time span and asset class. In particular, we demonstrate the usefulness of the proposed modeling framework by comparing the full sample parameter estimates and out-of-sample forecasting performance of specific RLS-ARFIMA models relative to that from other popular models in the literature.

The full sample parameter estimates reveal that random level shifts are important components of all series and that a genuine long memory component is present in the volatility series constructed using high-frequency data. The remaining dynamics in volatility proxies constructed as log-daily absolute returns, on the other hand, may be described as a combination of short memory dynamics and measurement errors. Finally, we show that the RLS-ARFIMA model display consistently good out-of-sample performance across forecast periods, forecast horizons, asset classes, and volatility measures, by being the most frequent model in the 10% MCS of Hansen et al. (2011). The forecast gains can be very pronounced at longer horizons. This shows that there is substantial statistical value in distinguishing between random level shifts and genuine long memory for forecasting.

References


respectively, using a bandwidth $\lfloor T^{1/2} \rfloor$.

Table 2: Simulation results for an RLS-LMSV(1, $d$) process fitted to an RLS-LMSV(1, $d$) process with parameters $d = 0.35$, $\gamma/T = 0.02$, $\sigma_\eta = 0.5$, $\sigma_\epsilon = 3\sigma_\eta$, $\phi = 0.2$ and two different levels of measurement errors specified through the noise-to-signal ratio, $\xi = \sigma_\epsilon^2/\sigma_\eta^2(1 - \phi)^2$, where we, specifically, consider $\xi = (1, 2)$. Furthermore, we vary the sample size $T = (3000, 6000)$, truncations $M = (20, 30, T^{1/2})$, and consider $N = 100$ replications.
<table>
<thead>
<tr>
<th>RLS-LMSV(1, d)</th>
<th>( \phi )</th>
<th>( d )</th>
<th>( \gamma/T )</th>
<th>( \sigma_\eta )</th>
<th>( \sigma_\epsilon )</th>
<th>( \sigma_u )</th>
<th>KMLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAC HF</td>
<td>-0.0329</td>
<td>0.4795</td>
<td>0.0169</td>
<td>0.2721</td>
<td>0.2248</td>
<td>0.0004</td>
<td>135.934</td>
</tr>
<tr>
<td>MRK HF</td>
<td>0.0470</td>
<td>0.3063</td>
<td>0.0152</td>
<td>0.8466</td>
<td>0.3063</td>
<td>0.0000</td>
<td>-188.044</td>
</tr>
<tr>
<td>SPY HF</td>
<td>0.0063</td>
<td>0.4106</td>
<td>0.0193</td>
<td>0.3794</td>
<td>0.2266</td>
<td>0.0000</td>
<td>57.2792</td>
</tr>
<tr>
<td>S&amp;P 500 HF</td>
<td>-0.0266</td>
<td>0.3168</td>
<td>0.0263</td>
<td>0.8738</td>
<td>0.2602</td>
<td>0.0891</td>
<td>-1620.01</td>
</tr>
<tr>
<td>T-Bonds HF</td>
<td>0.0254</td>
<td>0.4936</td>
<td>0.0051</td>
<td>0.3275</td>
<td>0.1355</td>
<td>0.3294</td>
<td>-2902.42</td>
</tr>
<tr>
<td>USD-AUD</td>
<td>0.0146</td>
<td>0.0277</td>
<td>0.0028</td>
<td>1.0459</td>
<td>1.3567</td>
<td>0.2651</td>
<td>-16933.4</td>
</tr>
<tr>
<td>USD-CHF</td>
<td>-0.0492</td>
<td>0.0874</td>
<td>0.0017</td>
<td>0.6769</td>
<td>1.2413</td>
<td>0.1638</td>
<td>-15872.1</td>
</tr>
<tr>
<td>USD-YEN</td>
<td>0.6903</td>
<td>0.0000</td>
<td>0.0028</td>
<td>2.9687</td>
<td>0.2516</td>
<td>1.2301</td>
<td>-16297.7</td>
</tr>
</tbody>
</table>

Table 3: Parameter estimates of the RLS-LMSV(1, d) model for the eight log-volatility series. “KMLE” denotes the predictive log-likelihood value from the Kalman filter.
Table 4: Simulation results using the following configurations: RLS-ARFIMA$(1,d,1)$ with $\phi = 0.2$, $\theta = -0.1$, $\sigma_\epsilon = 0.5$ and (DGP 1) $d = 0$, $\gamma/T = 0.02$, $\sigma_\eta = 3\sigma_\epsilon$; (DGP 2) $d = 0.35$, $\gamma/T = 0$, $\sigma_\eta = 0$; (DGP 3) $d = 0.35$, $\gamma/T = 0.02$, $\sigma_\eta = 3\sigma_\epsilon$; (DGP 4) $d = 0.6$, $\gamma/T = 0.02$, $\sigma_\eta = 3\sigma_\epsilon$. The bias and root mean squared error (RMSE) are computed for different values of $M$ (the last entry for RLS-ARFIMA), $T = 3000$ and $N = 100$ replications.
Table 5: Parameter estimates of the various dynamic models with standard errors in parentheses for the high-frequency log-volatility proxies on BAC, MRK, SPY and S&P 500. “KMLE” denotes the predictive log-likelihood value from the Kalman filter. Here, $a$ refers to the constant in an ARFIMA model. The standard errors are computed using the (inverse) numerical Hessian matrix.
### Table 6: Parameter estimates of the various dynamic models with standard errors in parentheses for the log-volatility proxies on the T-bonds, USD-AUD, USD-CHF and USD-YEN. “KMLE” denotes the predictive log-likelihood value from the Kalman filter. Here, \( a \) refers to the constant in an ARFIMA model. The standard errors are computed using the (inverse) numerical Hessian matrix.
Table 7: Forecast evaluations of the eight dynamic models. We use mean squared forecast errors (MSFE's) and consider MCS comparisons with all models included in the initial set. Here, boldface notation indicate whether a model belongs to the 10% MCS. The MCS $p$-values are in parenthesis. See the main text for details.
<table>
<thead>
<tr>
<th>Model</th>
<th>$t_{out} \in [1, 300]$</th>
<th>$t_{out} \in [301, 600]$</th>
<th>$t_{out} \in [601, 900]$</th>
<th>$t_{out} \in [1, 900]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RLS-ARFIMA(0, d, 0)</td>
<td>1.2418</td>
<td>5.9518</td>
<td>13.781</td>
<td>1.7523</td>
</tr>
<tr>
<td>RLS-ARFIMA(1, d, 1)</td>
<td>1.2273</td>
<td>5.1191</td>
<td>10.526</td>
<td>1.6613</td>
</tr>
<tr>
<td>RLS-ARMA(0, 0)</td>
<td>1.2329</td>
<td>6.1010</td>
<td>14.548</td>
<td>1.7596</td>
</tr>
<tr>
<td>RLS-ARMA(1, 1)</td>
<td>1.2401</td>
<td>5.9229</td>
<td>13.666</td>
<td>1.7472</td>
</tr>
<tr>
<td>ARFIMA(0, d, 0)</td>
<td>1.3017</td>
<td>9.2743</td>
<td>30.938</td>
<td>1.7086</td>
</tr>
<tr>
<td>ARFIMA(1, d, 1)</td>
<td>1.2517</td>
<td>42.195</td>
<td>199.33</td>
<td>1.6895</td>
</tr>
<tr>
<td>HAR</td>
<td>1.2624</td>
<td>9.4700</td>
<td>53.660</td>
<td>1.6905</td>
</tr>
<tr>
<td>log-GARCH</td>
<td>1.9869</td>
<td>30.249</td>
<td>126.68</td>
<td>2.5164</td>
</tr>
<tr>
<td></td>
<td>$t_{out} \in [601, 900]$</td>
<td>$t_{out} \in [1, 900]$</td>
<td>$t_{out} \in [601, 900]$</td>
<td>$t_{out} \in [1, 900]$</td>
</tr>
<tr>
<td>RLS-ARFIMA(0, d, 0)</td>
<td>1.3759</td>
<td>11.051</td>
<td>31.781</td>
<td>1.4566</td>
</tr>
<tr>
<td>RLS-ARFIMA(1, d, 1)</td>
<td>1.3155</td>
<td>10.272</td>
<td>28.001</td>
<td>1.4013</td>
</tr>
<tr>
<td>RLS-ARMA(1, 1)</td>
<td>1.3770</td>
<td>11.102</td>
<td>32.006</td>
<td>1.4548</td>
</tr>
<tr>
<td>ARFIMA(0, d, 0)</td>
<td>1.3681</td>
<td>30.555</td>
<td>128.80</td>
<td>1.4594</td>
</tr>
<tr>
<td>ARFIMA(1, d, 1)</td>
<td>1.3224</td>
<td>82.939</td>
<td>398.10</td>
<td>1.4212</td>
</tr>
<tr>
<td>HAR</td>
<td>1.3247</td>
<td>16.070</td>
<td>93.273</td>
<td>1.4258</td>
</tr>
</tbody>
</table>

Table 8: Forecast evaluations of the eight dynamic models. We use mean squared forecast errors (MSFE's) and consider MCS comparisons with all models included in the initial set. Here, **boldface** notation indicate whether a model belongs to the 10% MCS. The MCS $p$-values are in parenthesis. See the main text for details.
Table 9: Forecast evaluations of the eight dynamic models. We use mean squared forecast errors (MSE's) and consider MCS comparisons with all models included in the initial set. Here, **boldface** notation indicate whether a model belongs to the 10% MCS. The MCS p-values are in parenthesis. See the main text for details.
Out-of-Sample Series: T-Bonds HF

Figure 1: The upper panel displays the cumulative ten-step-ahead log-volatility proxy for the T-bond series. The middle and lower panels display the corresponding loss differentials, $d_{ij,t} = (\bar{\sigma}_{t,\tau} - \bar{y}_{t+\tau,i})^2 - (\bar{\sigma}_{t,\tau} - \bar{y}_{t+\tau,j})^2$, from the comparisons of the RLS-ARFIMA(0, d, 0) model against the RLS-ARFIMA(1, d, 1) model and the RLS-ARMA(0, 0) model against the RLS-ARMA(1, 1) model.